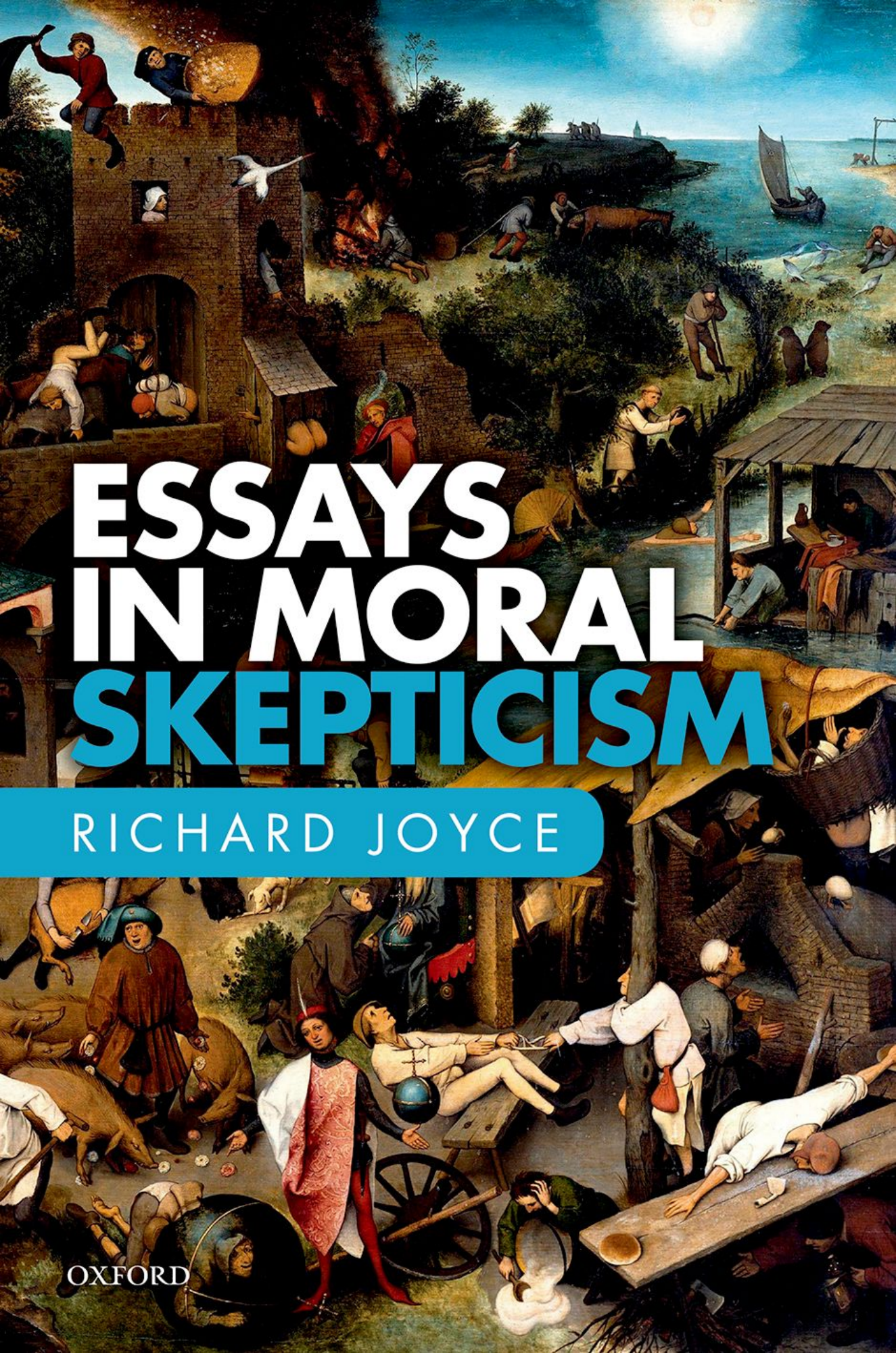
ESSAYS
IN MORAL
SKEPTICISM
RICHARD JOYCE
OXFORD

Essays in Moral Skepticism

Essays in Moral Skepticism

Richard Joyce

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

© in this volume Richard Joyce 2016

The moral rights of the author have been asserted

First Edition published in 2016
Impression: 1

All rights reserved. No part of this publication may be reproduced, stored in
a retrieval system, or transmitted, in any form or by any means, without the
prior permission in writing of Oxford University Press, or as expressly permitted
by law, by licence or under terms agreed with the appropriate reprographics
rights organization. Enquiries concerning reproduction outside the scope of the
above should be sent to the Rights Department, Oxford University Press, at the
address above

You must not circulate this work in any other form
and you must impose this same condition on any acquirer

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data
Data available

Library of Congress Control Number: 2015944573

ISBN 978–0–19–875487–9

Printed in Great Britain by
Clays Ltd, St Ives plc

Links to third party websites are provided by Oxford in good faith and
for information only. Oxford disclaims any responsibility for the materials
contained in any third party website referenced in this work.

For Max and Lucia

Contents

Acknowledgments ix

Introduction: Morality: The Evolution of a Myth 1

Part I. Error Theory

1. Expressivism, Motivation Internalism, and Hume 17
2. Morality, Schmorality 41
3. The Accidental Error Theorist 67
4. Metaethical Pluralism: How both Moral Naturalism and Moral Skepticism may be Permissible Positions 89

Part II. Evolution and Debunking

5. The Origins of Moral Judgment 109
6. The Many Moral Nativisms 122
7. Evolution, Truth-Tracking, and Moral Skepticism 142
8. Irrealism and the Genealogy of Morals 159

Part III. Projectivism and Fictionalism

9. Patterns of Objectification 177
10. Is Moral Projectivism Empirically Tractable? 195
11. Moral Fictionalism 219
12. Psychological Fictionalism, and the Threat of Fictionalist Suicide 240

References 257
Index 273

Acknowledgments

The essays in this volume are reprinted by permission of the original publishers:

"Expressivism, motivation internalism, and Hume." In C. Pigden (ed.), *Hume on Motivation and Virtue* (Palgrave MacMillan, 2010).

"Morality, schmorality." In P. Bloomfield (ed.), *Morality and Self-Interest* (Oxford University Press, 2007).

"The accidental error theorist." In R. Shafer-Landau (ed.), *Oxford Studies in Metaethics, Vol. 6* (Oxford University Press, 2011).

"Metaethical pluralism: How both moral naturalism and moral skepticism may be permissible positions." In S. Nuccetelli and G. Seay (eds.), *Ethical Naturalism: Current Debates* (Cambridge University Press, 2012).

"The origins of moral judgment." *Behaviour* 151, special issue: *Evolved Morality: The Biology and Philosophy of Human Conscience* (2014).

"The many moral nativisms." In K. Sterelny, R. Joyce, B. Calcott, and B. Fraser (eds.), *Cooperation and its Evolution* (MIT Press, 2013).

"Irrealism and the genealogy of morals." *Ratio* 26, special issue: *Irrealism in Ethics* (2013).

"Patterns of objectification." In R. Joyce and S. Kirchin (eds.), *A World Without Values: Essays on John Mackie's Moral Error Theory* (Springer Press, 2010).

"Is moral projectivism empirically tractable?" *Ethical Theory and Moral Practice* 12 (2009).

"Moral fictionalism." In M. Kalderon (ed.), *Fictionalism in Metaphysics* (Oxford University Press, 2005).

"Psychological fictionalism, and the threat of fictionalist suicide." *The Monist* 96 (2013).

Introduction

Morality: The Evolution of a Myth

People will often express their strong moral convictions as claims of *knowledge*. Surely we know that making a sport of torturing innocents is evil, know that sympathy is morally better than spite, know that defrauding thousands of people of their retirement savings for no motive other than greed is morally deplorable? One would likely be confused by someone's claiming, seemingly sincerely, not to know such things. As when faced with somebody who honestly claims not to know whether kangaroos are animals, or claims not to know whether three is a number, one would probably be baffled as to what breakdown might lie behind such a fundamental epistemic flaw, and in all likelihood would feel unsure where even to begin correcting such a person. More than this, in the moral case we may not even feel that epistemic correction is quite the appropriate course. G. E. M. Anscombe (1958: 17) writes of someone who sincerely claims not to know that executing innocents is morally wrong: "I do not want to argue with him; he shows a corrupt mind."

Despite the importance of the idea of moral knowledge in human affairs, there is a very long philosophical tradition of doubting that any such thing exists. Sometimes this moral skepticism falls out of a more general epistemological skepticism; sometimes it is specifically *moral*. If we accept the mundane (though far from incontestable) view that to know that *p* involves (i) believing that *p* (ii) truly, and (iii) with justification, then moral skepticism is the disjunction of three theses:

a) Noncognitivism: the denial that moral judgments express beliefs.
b) Error theory: the acceptance that moral judgments express beliefs, but the denial that moral judgments are ever true.
c) Justification skepticism: the acceptance that moral judgments express beliefs, but the denial that moral judgments are ever justified.

While both (b) and (c) are explicitly contraries of (a), they are not contraries of each other. One might endorse an error theory while maintaining that people are justified in their moral beliefs, or alternatively endorse an error theory while adding that all people's moral beliefs lack justification. Similarly, the claim that moral beliefs lack justification may combine with the view that they are all false, but is also consistent with the possibility that moral beliefs are not only true but objectively true. I add this last point

about objectivity in order to draw attention to the fact that moral skepticism need not be construed as a form of moral anti-realism. If we take moral realism to be the view that moral discourse expresses beliefs that are sometimes true and, when true, are true in virtue of the obtaining of objective facts (under some specification of *objectivity*), then justification skepticism is compatible with a realist stance. Conversely, if one maintains that moral discourse expresses beliefs about some realm of *non-objective* facts—beliefs that are often both true and justified—then one will be a moral anti-realist but not a moral skeptic.

It's a good thing for me that (b) and (c) are not contraries, since some time ago I wrote a book devoted largely to arguing for (b)—*The Myth of Morality* (2001)—which I followed up a few years later with a book devoted largely to arguing for (c)—*The Evolution of Morality* (2006). While I purposely gave these books titles that appear in tension (a decision that can be put down to nothing more than perverse philosophical misbehavior[1]), in terms of content they were intended to be consistent with each other. *The Myth of Morality* argues that all moral claims are (though expressions of belief) untrue, and it does so using ordinary metaethical methods. *The Evolution of Morality* argues that all moral claims are (though expressions of belief) unjustified, and it does so using an argument that includes an empirical appeal to the evolutionary origins of human moral thinking (also known as an "evolutionary debunking argument"). The two books' central arguments don't interact much logically, and their respective skeptical conclusions are independent of each other—though, thankfully, compatible.

The claim that either of these conclusions implies the falsity of noncognitivism requires some qualification. As I've just presented it, noncognitivism is an entirely negative thesis. As a matter of fact, however, it is always (so far as I know) also put forward as a positive thesis: "Moral judgments do not express beliefs, but rather they linguistically function to do so-and-so" (where the openness of the "so-and-so" is what accounts for different forms of noncognitivism). It is natural to read this as asserting that "doing so-and-so" (where this is something other than expressing beliefs) is the *only* linguistic function of moral judgments. Let us call such a view "pure noncognitivism." A weaker noncognitivist perspective places the "only" elsewhere in the sentence: "Moral judgments do not *only* express beliefs, but rather they [also] linguistically function to do so-and-so." This softened view recognizes that moral judgments may have complex linguistic functions in a way that the pure view does not. While (b) and (c) imply the falsity of pure noncognitivism, they do not imply the falsity of weaker varietals. (This is discussed in Essay 1 of this collection.) In other words, one needs to make theoretical space for a hybrid view which claims that moral judgments both express beliefs *and* perform noncognitive function so-and-so. The beliefs in question may be true in virtue of objective facts (making for a kind of realist–noncognitivist mix), or may be

[1] The title of the introduction that you are currently reading may be seen as some sort of belated Hegelian resolution of the two (not that I flatter myself in thinking that anyone has been waiting with bated breath!).

never true (making for an error-theoretic–noncognitivist mix), or may be unjustified (making for a mix of noncognitivism and justification skepticism). If, therefore, moral skepticism is the disjunction of (a), (b), and (c), then the most extreme skeptical view available will combine elements of all three disjuncts: Moral judgments express beliefs but they also perform noncognitive functions, and the beliefs in question are both false and unjustified. I am inclined to accept this extreme view.

This collection traces out the broad strokes of my main metaethical preoccupations and the development of my views since publishing the two aforementioned books. It is divided, somewhat imperfectly, into three parts of four essays each. The essays of Part I ("Error Theory") follow on most directly from *The Myth of Morality*; the essays of Part II ("Evolution and Debunking") upgrade ideas presented in *The Evolution of Morality*; and Part III ("Projectivism and Fictionalism") has the twin focal points of two related theses that were discussed in the two books but for which I felt there was more to be said.

I stand by the earlier skeptical spirit, but in some ways my views have shifted. Sometimes these are relatively minor adjustments, responses to criticisms, the straightening out of confusions (both mine and others'), or the application of old ideas to new areas. Two of the essays herein (4 and 6) articulate what are in principle quite major revisions in view: where the possibility of conceptual indeterminacy leaves matters that I once thought decidable in the skeptic's favor potentially undecidable. In these situations I plump for a kind of ecumenical pluralism. Whether one sees this as renouncing the idealistic tendencies of youth, or as coming to appreciate a more nuanced metaethical Big Picture, is (in a phrase from David Lewis that echoes through the papers in question) "mainly a matter of temperament."

In what follows of this Introduction I will present an overview of each section, though I will not make an especial effort to describe every essay in turn; generally, they speak for themselves.

Error Theory

To take an error-theoretic stance toward a discourse is to maintain that the discourse consists of assertions that fail to be true. A paradigm familiar example is the atheist's attitude toward religion. It seems reasonable to hold that most religious utterances are expressions of the speaker's beliefs (for example, the belief that the gods care whether we keep or break our promises), and the atheist is confident that the world is not furnished with the objects/properties/relations necessary to render these beliefs true (for example, there exist no such caring gods).

A natural way of interpreting the atheist is as holding that religious concepts are reasonably well-defined but that the world contains nothing answering to these definitions, yet an atheist might also maintain that the fault lies with the religious concepts (*God*, *karma*, *sin*, and so on) being in some manner hopelessly confused (though still

able to appear in assertions[2]). Similarly, moral error theorists might be moved by various kinds of arguments. One kind of error theorist will seek first to pin down the meaning of moral concepts (*obligation*, *evil*, *moral goodness*, and so on) and then argue that the world lacks any properties satisfying these concepts. Another kind of error theorist might admit bewilderment as to what these moral concepts are supposed to denote; she might say to moral believers: "Look, you can't even seem to agree among yourselves as to what the central moral concepts denote, and, moreover, your disagreement is sufficiently deep-seated that I'm left doubting whether these concepts in common usage even denote anything at all." Just as there are both atheists who believe that the non-existence of gods is necessary and atheists who believe that it is contingent, so too should the label "moral error theorist" cover both those who maintain that moral properties exist at no possible world and those who maintain that their non-existence is merely contingent.

The principle alternatives to the error-theoretic view are noncognitivism, moral realism, and (for want of a better title) moral non-objectivism. The last is the view that our moral discourse succeeds in referring to moral properties of a non-objective nature.[3] (Think, by analogy, of what it takes for something to be *illegal*. It is certainly a fact that it is illegal to drive on the right-hand side of the road in New Zealand, but it is a fact constituted by our collective decisions/beliefs/practices—it is not, in the relevant manner, an *objective* fact.) The error-theoretic view can be defined in terms of what it agrees with and what it rejects from these alternatives. The error theorist agrees with the realist and the non-objectivist (and not with the pure noncognitivist) that moral discourse functions to express our beliefs about the moral status of various aspects of the world. A standard argument for the error theory (defended by John Mackie and myself) agrees with certain moral realist views (and not with non-objectivist views) about what moral facts would have to be like in order for our moral judgments to be true. Mackie, for example, thinks that a sort of Kantian/Moorean moral realism gets things basically correct at the conceptual level. But the error theorist parts company from the realist (and joins company with the noncognitivist and the non-objectivist) in maintaining that the realistic conception of morality asks too much of the world; there is nothing answering to the Kantian/Moorean conception of moral facts, for example.

Defining the error-theoretic position in this fashion allows us to see that it contains no unique or outrageous sub-thesis; at each step, taken in isolation, one should expect to find many non-error-theorists nodding enthusiastically and offering their own arguments in support. Indeed, the error theorist could get by without developing any

[2] I realize that this parenthetical qualification raises some substantive questions; unfortunately I lack space to pursue them here.

[3] Earlier in this Introduction, "moral realism" was defined as the conjunction of three theses: belief, truth, and objectivity. Moral non-objectivism is the endorsement of the first two and the denial of the third. Since the *objectivity/non-objectivity* distinction is notoriously difficult to articulate (see Essay 10), this taxonomy is a slippery one.

novel argument of her own, simply by purloining all her arguments from other metaethical positions. What the error theorist does is combine these arguments and subtheses in a manner that leads to a radical skepticism that many find unpalatable and threatening (and, perhaps, wishing to retract their supportive arguments!).

In *The Myth of Morality* I raised some fairy blunt considerations against noncognitivism. In Essay 1 of this collection, "Expressivism, Motivation Internalism, and Hume," I revisit the issue more conscientiously, advocating a weak form of noncognitivism but remaining firm against pure noncognitivism, and careful to clarify how this concession steals no wind from the error theorist's sails. (I also offer an interpretation of Hume along these lines, though it has a decided "for-what-it's-worth" air.[4]) It is not, however, the error theorist's take on the dispute between the noncognitivist and the cognitivist on which I wish to focus here, but rather the error theorist's take on the dispute between the realist and the non-objectivist.

Earlier I spoke of a kind of error theorist who tries to pin down the meaning of moral concepts (*obligation*, *evil*, *moral goodness*, and so on), and I subsequently referred to a "standard argument" for the error theory being one that agrees with certain realists regarding the conceptual commitments of moral discourse. Before proceeding, let me stress that endorsing such a view is not a necessary feature of the error-theoretic position; it is, rather, *a strategy for arguing for* that position. There are other possible grounds for becoming a moral error theorist. One might, for example, come to accept the moral error theory through becoming thoroughly disillusioned with all other metaethical positions. Or alternatively consider, for example, a non-objectivist view with the simple structure "Moral goodness = Nness" (where "Nness" denotes some naturalistic non-objective property). The standard error-theoretic complaint against such a theory is that Nness lacks the distinctive practical authority with which moral properties are essentially imbued, and therefore the non-objectivist's equation can be rejected by appeal to Leibniz's law. But instead the error theorist might embrace the non-objectivist's equation yet argue that it nevertheless leads to an error theory because "Nness" fails to denote any actually instantiated property. I explore this alternative (non-standard) strategy for the error-theoretic conclusion in Essay 3 of this collection, "The Accidental Error Theorist." This non-standard argument has limits, of course, for it can hardly be argued that "Nness" *must* suffer from this failure. For those many occasions where the non-objectivist's "Nness" succeeds in picking out an actually instantiated property, the error theorist must return to the standard strategy of rejecting the reasonableness of the equation.

One can consider the standard strategy either generally or specifically. Speaking generally, the strategy identifies some thesis to which moral discourse is committed and

[4] I'm no Hume scholar. My "serious" attempts at history of philosophy came early in my career—"Early Stoicism and akrasia" (*Phronesis* 1995) and "Cartesian memory" (*Journal of the History of Philosophy* 1997)—efforts aroused by wonderful teachers at Princeton: John Cooper and Margaret Wilson. I was quite tempted to write a PhD dissertation under Margaret's supervision on early modern conceptions of secondary qualities, but at the eleventh hour veered into metaethics (with Gilbert Harman).

then argues that the thesis is false. Thus this argumentative strategy faces two kinds of opponent: Those who reject the error theorist's conceptual step engage in a semantic dispute, while those who reject the latter step disagree about what features the world contains. The semantic dispute is, I think, the trickier to prosecute, for it is challenging to know how best to articulate the difference between a discourse being *committed* to some thesis (in the sense that dropping that thesis would amount to changing the subject) and a discourse being such that people sometimes/often/always have some false beliefs concerning it. Considering the strategy in general terms leaves open what the specific problematic thesis (or theses) might be. Speaking more specifically, perhaps moral discourse is committed to a problematic notion of desert, or a problematic notion of autonomy, or a problematic notion of personhood, or a problematic epistemology. The specific version of the argument that has dominated discussion, however, is that moral discourse is committed to a problematic notion of objective practical authority.

I don't mind confessing that I've never really nailed the conceptual step of this argument to my own satisfaction; but, on the other hand, I've never found the efforts of those opposed to the step terribly persuasive either. Part of the challenge is to render the idea of "objective practical authority" in a sufficiently clear manner, and part of the challenge is to establish that moral discourse is committed to such a thing. Regarding the former, although in *The Myth of Morality* I gave it my best shot—hypothesizing that this authority might be understood by reference to certain kinds of practical reasons—even then I felt that it may be asking too much of the error theorist to provide this much specificity. After all, error theorists may worry that there is something utterly mysterious about the kind of authority with which moral properties are essentially imbued; they may consider morality to be something like a pseudoscience, and the concepts (or pseudo-concepts) employed by a pseudoscience often defy clarification. By analogy: As an atheist I don't believe that anything is literally *sacred*, yet I don't suppose I could do an especially good job of articulating precisely what it means for something to be sacred; it is, rather, the very obscure and nebulous quality of the concept that encourages my disbelief. The fact that the atheist may be unable to draw a very precise bead on concepts like *sacred*, *God*, or *heaven* shouldn't count against the reasonableness of his atheism; it hardly seems incumbent upon *him* to give definition to these ideas! That said, of course the atheist needs to have *some* idea of the content of these concepts, or else he could not object to someone who tries to reassure him that "sacred" means nothing more than *salubrious*, that "God" just means *love*, and that "heaven" is a word that denotes *Tahiti*. (And how can one reasonably doubt the existence of salubrity, love, and Tahiti?) The atheist needs to be sufficiently conceptually *au fait* to protest that these religious concepts are used to denote something other than such innocuous entities. In a similar way, when faced with a moral naturalist who proposes to identify moral properties with some kind of innocuous naturalistic property—the maximization of happiness, say—the error theorist will likely object that this property lacks the kind of "normative oomph" that permeates our moral discourse. Why, it might be asked, should we care about the maximization of happiness any more

than the maximization of some other mental state, such as surprise?[5] Yet (the error theorist may continue) moral properties are those about whom the failure to care counts as a transgression; this is in fact the *whole point* of having a moral discourse. The error theorist's defense here relies on identifying a conceptual commitment of morality, though not necessarily a precisely defined one.

There has been some work in recent years trying to ascertain the conceptual commitments of moral discourse via empirical methods. (See Goodwin and Darley 2008, 2012; Sarkissian et al. 2011; Uttich et al. 2014.) A typical experimental design is to have test subjects assess various kinds of normative claims (such as "Wearing pajamas and a bathrobe to a seminar is wrong behavior") as true or false or a matter of opinion. Of course, even if all subjects are adamant that moral norms have quality Q it would not follow that Q is an indispensible conceptual commitment of moral discourse, but one might be tempted to conclude that such a result would at least show that *not-Q* is not such a commitment. This temptation should be treated with care. What students tick on a questionnaire need not reflect their real moral commitments; these emerge only in the testing ground of actual practice. Such experiments are also prone to over-interpretation or misinterpretation. For example, all the experiments that I have seen along these lines seem to target the extent to which subjects make pronouncements in line with a *relativistic* metaethical view, yet the conclusion drawn over and over again is that this has some bearing on the subjects' attitudes toward moral *objectivity*. But any metaethicist worth his or her salt will tell you that the opposite of relativism is *absolutism*, not objectivism. Relativism-versus-absolutism doesn't figure in the criteria distinguishing moral realism from non-realism.[6] The proposal that moral discourse is imbued with a kind of objective practical authority (for want of a better phrase) is entirely consistent with a relativistic moral discourse.

The seeming impasse between the error theorist and her critics over what is and is not a conceptual commitment of moral discourse should put one in a diagnostic mood. How can thoughtful and intelligent people disagree over the content of their own concepts? I have become increasingly sympathetic to the idea that the line between a discourse having a faulty conceptual commitment and a discourse being such that users tend to have false beliefs about its subject matter is an extremely blurry one—not just in an epistemic sense, but in the sense that there often is no fact of the matter. The moral concepts are indeterminate beasts, available for different equally legitimate but non-equivalent precisifications. Some such precisifications may find something answering to them in the world, while others of the same concept may not. In other words, certain forms of moral naturalism may be permissible to maintain, but the moral error theory may also be permissible to maintain. This viewpoint, and the resulting metaethical pluralism, is the subject of Essay 4 of this collection. In this Essay

[5] I should quite like to see someone defend *thaumatistic utilitarianism*: the thesis that one is obligated to maximize surprise. I suspect that the silliness of the enterprise would cast some light on forms of utilitarianism that are taken seriously, like hedonic utilitarianism.

[6] See Joyce 2007a.

I also make a plea for the pragmatic value of the error-theoretic position; I do this to provide a counterweight to what I suspect will be the widespread assumption that this kind of Scottish verdict does not really encourage pluralism but rather plays into the hands of the moral naturalist. Here we get to some really rather deep issues about what the point of philosophy is supposed to be: Should a theory that vindicates the vernacular be preferred to one that does not? I am yet to encounter an argument that convinces me that a positive answer to this question is mandatory; rather, I feel more inclined to side with the bleak romanticism expressed by Edward Gorey: "My mission in life is to make everybody as uneasy as possible. I think we should all be as uneasy as possible, because that's what the world is like." Well, maybe not as uneasy *as possible*, but I certainly dislike the image of philosophy as a tool for providing a soothing background voice reporting the world to be as we believe it to be, moral facts and all.

Evolution and Debunking

The atheist accuses the vast majority of human beings of embracing a doxastic error of enormous proportions. Yet if asked "Where does this error come from?" I don't think the atheist's position is particularly undermined if she admits that she has no idea beyond, perhaps, having a somewhat pessimistic view of humans as silly and gullible creatures. In the same way, I don't judge it incumbent on the moral error theorist to offer a theory of why nearly all humans have fallen into the mistaken ways of moral thinking. Yet it also seems reasonable to claim that both the atheist and the moral error theorist would strengthen their positions somewhat if each could provide a plausible hypothesis concerning how such systematic errors might arise in human thought. In *The Myth of Morality* (chapter 6) I turned to evolution as a plausible explanation of human moral thinking, observing that if Darwinian selection has wired the human brain for moral judgment, it is because moral judgment enhanced our ancestors' reproductive fitness (relative to competitors) in a way that might be accomplished even by encouraging false beliefs.

While developing these thoughts I became intrigued by another possibility: that one might argue for a moral error theory *on the basis of* these evolutionary considerations. My next book, *The Evolution of Morality*, was intended to investigate this argument comprehensively, but in the end a great many other interesting things arose and only the final chapter was devoted to the debunking argument (as it has become known). By the time I wrote this final chapter it had also become clear to me that the error-theoretic conclusion is beyond the reach of the argument; the correct skeptical conclusion is of an epistemological nature: that all moral judgments are *unjustified*.[7] Thus *The Evolution of Morality* does not argue for moral anti-realism at all.

[7] In a manner very much bringing to mind Bertrand Russell's comparison of theft and honest toil, I decided that one might nevertheless call this epistemological conclusion a version of "error theory" (Joyce 2006a: 223). This was, I now see, a foolish stipulation, and I discourage anyone from adopting this usage.

At this time Sharon Street was writing an influential paper, "A Darwinian dilemma for realist theories of value," whose argument starts with the same premise as mine: that the human tendency to assess the world in moral terms is a biological adaptation. (Neither of us, of course, purport to be in a position to assert this moral nativist thesis with great confidence; our arguments are conditional.) From this similar starting point, though, our arguments diverge importantly. Street's target is moral realism; she uses evolutionary considerations to cast doubt on the existence of objective moral facts. Since, however, she is willing to accept a constructivist metaethical view—according to which moral facts have a non-objective status—she is no error theorist. The reason I mention Street's argument at this point is that it appears to me that because our arguments came out at much the same time there has been a tendency to lump them together. A cottage industry focused on "evolutionary debunking arguments of morality" has sprouted (most of it critical), and while much of it is very worthwhile, some of it is also, in my opinion, based on fundamental misunderstandings. I take some blame for this, since my attempt to make the case in *The Evolution of Morality* is flawed and unclear in various ways; I was still squinting to discern the structure of my own argument. Essays 7 and 8 of this collection represent my attempts to develop the argument more cleanly. (The ordering of this pair of essays reflects the sequence in which they were written, but whereas Essay 8 was published in 2013, the project for which Essay 7 was originally commissioned was delayed and remains uncertain; hence Essay 7 is published here for the first time.)

As I mentioned, the debunking argument is conditional: It relies on an empirical premise concerning the evolution of morality which is yet to be established. Before we come to the debunking argument, then, it makes sense to examine this nativist hypothesis carefully. This is done in Essays 5 and 6 of this collection. Here metaethics is put mostly to one side, and instead I adopt my role as a philosopher of biology—albeit, I'll be the first to admit, as something of a Sunday painter.

One of the flaws of *The Evolution of Morality* is that I hadn't yet come to appreciate fully the difficulties surrounding the key notion of *innateness*; my subsequent papers on the topic are more cognizant of this. Essay 5, "The Origins of Moral Judgment," focuses on the question of how we are to distinguish traits that are adaptations from those that are byproducts; Essay 6, "The Many Moral Nativisms," looks at different meanings of "innateness." Both papers are to some extent motivated by desire to respond to a moral anti-nativist movement that emerged in the years after my book. It's not so much that I think that the anti-nativists (or "spandrel theorists" as I sometimes call them) are definitely mistaken; I am more interested in diagnosing the conceptual framework of the disagreement and straightening out misunderstandings.

Something to which all should agree is that there's no point in arguing over whether some trait is or is not innate (regardless of what notion of *innateness* is under discussion) unless we have a reasonable grasp of what the trait is that we're arguing about. In the present case, that trait is the capacity to make moral judgments. So: What is it to make a moral judgment? At one time I was keen to argue for a particular kind of

answer, but years of encountering critics with very different views have led me to suspect that there might be no fixed answer; the matter is simply indeterminate in various ways. This suspicion, obviously, matches that voiced in Essay 4. There the worry was that our moral concepts are indeterminate, so that the judgment "ϕ is morally prohibited" (say) may be reasonably considered true or reasonably considered false, depending on how one precisifies the idea of *moral prohibition*: Understood one way it picks out an actual property; understood another way it fails to do so. Here the worry is that what it takes to make a moral judgment suffers from indeterminacy, so that the sentence "S judges that ϕ is morally prohibited" (say) may be reasonably considered true or reasonably considered false, depending on how one precisifies the idea of making this kind of moral judgment. The node of indeterminacy examined in Essay 6 is that one understanding of *moral judgment* constructs the phenomenon entirely out of noncognitive building blocks, whereas another requires of it more conceptual sophistication. The possibility then arises that understanding the phenomenon one way may pick out a capacity that is innate, but understanding it another way may pick out a capacity that is not innate; hence, the debate between the moral nativist and the anti-nativist may be undecidable.[8]

Projectivism and Fictionalism

A question that has always divided philosophers (if one may be so crass as to speak of such a thing) is how close the world really is to how it seems to us. Those of a skeptical temperament, who lean toward thinking that it may not be very close at all, have always faced the further question of how one should respond, in practical and psychological terms, to this human condition. An ancient response—as ancient, at least, as Pyrrho and his followers—is that it shouldn't make too much difference at all; we can and should carry on living in accordance with appearances. Projectivism is a way of making sense of the skeptical answer to the first question, and fictionalism is a way of making sense of the skeptical answer to the second question. Both theories rely on a similar distinction being drawn between how the world appears to us (/is experienced by us) and our critical understanding of what is really going on. Both theories also may have general applicability or can be restricted to the moral realm.

Moral projectivism is the view that our emotional life creates and colors our moral experience. One's seeing an act of violence as *wrong* (say) is not the result of successfully tracking the presence of wrongness, but is rather the result of having an emotion like disapproval which plays an active role in constructing how that action seems to us. The moral skeptic need not endorse projectivism, but doing so can be useful for the skeptic inasmuch as it can provide an explanation for how this systematic mistake in human thinking comes about. The skeptic can even argue that humans have been

[8] Please note the modal qualifications in my expression. In neither Essay 4 nor Essay 6 am I asserting that there *is* this indeterminacy; I am merely exploring the possibility in a sympathetic mood.

designed by natural selection to perform this kind of projective error—that the mechanisms underlying the process were adaptive for our ancestors—thus moral projectivism sits quite comfortably with the moral nativist program.

Part of my interest in writing about this topic was born of a frustration with a widespread view of moral projectivism holding it to be the exclusive province of the noncognitivist. I recall as a graduate student feeling anxious that I must have weird or mistaken ideas about what projectivism is; I heard people speaking of "projectivist semantics" and couldn't work out what this might mean. Around 2009 I wrote two papers that tried to straighten things out, at least to my own satisfaction; both attempt to clarify the relation between moral projectivism and various positions in metaethics. Essay 9, "Patterns of Objectification," takes its title from a phrase used by Mackie (who prefers "objectification" to "projectivism"), and investigates what role projectivism plays (or could play or should play) in his argument for a moral error theory. (In this respect, Essay 9 could easily plug into Part I of this collection.) Essay 10, "Is Moral Projectivism Empirically Tractable?," teases apart the various sub-theses of different potential versions of projectivism. Here I am keen to push past the metaphors ("spreading," "gilding," "projecting") to get at some literal theses which can then be (in principle) tested for truth. I continue to think that moral projectivism is a theory ripe for experimental investigation. This testable core of moral projectivism is, however, metaethically neutral.

Moral fictionalism is a proposal for how the error theorist might carry on.[9] There is no need to eliminate morality entirely from our thoughts and language (the fictionalist declares); we can maintain its use as a kind of functional fiction. The fictionalist doesn't propose that we maintain morality as a set of *beliefs* and *assertions* (for the advice "Carry on believing something that you believe to be false" is likely to prove problematic in various obvious ways), but rather in a fashion reminiscent of a kind of highly-played make-believe. Like all pieces of advice, the reasonableness of the fictionalist's proposal depends on the outcome of a cost-benefit analysis. Making a make-believe of morality is not, of course, going to produce the same costs and benefits as sincerely believing morality. Believing that ϕing is morally obligatory is likely to strengthen one's motivation to ϕ more robustly than make-believing that ϕing is obligatory—so on the assumption that ϕing is beneficial, moral belief is better than moral make-believe. But this is not the relevant comparison, since for the skeptic the option of belief is gone. Rather, we must compare the fictionalist proposal with that of the eliminativist, who counsels that we drop moral thinking and moral talk altogether. So long as thinking and talking of ϕing as morally obligatory (even in a fictionalist manner) *in some way* enhances one's motivation to perform the beneficial action (with no countervailing cost), then the fictionalist has made her case. (And it needn't be good advice

[9] Moral fictionalism can also be considered as a proposal for how a justification skeptic might carry on—though the way that the statement of the theory is worded would have to be adjusted in various ways. Think by analogy of religious fictionalism: The proposal might be offered both to atheists and agnostics. (Here "agnosticism" is used in the popular way to denote a position of doxastic indecision.)

for *all* the people *all* the time; that is not a requirement we ever put on something's counting as good advice!)

But how can mere pretense have a sufficient impact on one's motivations as to affect significant costs and benefits? The quick answer to this question (as to so many others) is that humans are strange creatures. And one of our stranger quirks is our interest in fictions. It is easy to imagine intelligent creatures for whom engaging with characters who never existed and narratives that never happened holds no attraction whatsoever. But that's not us. Stories and images affect our moods, emotions, and motivations. The error theorist who immerses herself in a moral fiction takes advantage of these peculiar aspects of her own psychology. She is likely acutely familiar with classifying her social environment in moral terms, so she continues the habit of bringing moral concepts to bear on practical problems, allowing moral emotions like disgust, anger, and guilt to wash through her. All going according to plan, this moral fiction doesn't encourage her to do anything that she wouldn't upon reflection choose to do anyhow on non-moral grounds. But moral thinking has some advantages over careful non-moral thinking, in that it can be fast and frugal, less prone to self-sabotaging rationalizations, able to banish practical calculation from the decision procedure when the very act of calculating is suboptimal. If the nativist is correct, then moral thinking is a well-honed tool that suits our psychological configuration; we are comfortable with its contours.

Essay 11, "Moral Fictionalism," develops the case I made in *The Myth of Morality*—struggling (as is always the case when I discuss this topic) to make a weird theory seem a bit less weird. (Some responses to critics of moral fictionalism also appear toward the end of Essay 2 of this collection.) I'm not sure whether my tentative advocacy of moral fictionalism over the years has won many converts, but in a sense this is how it should be, since the cost-benefit analysis upon which the theory rests involves so many unknown variables and counterfactuals that anyone who claims with confidence to believe that moral fictionalism is *correct* has probably missed the point.

The final essay of this collection, "Psychological Fictionalism, and the Threat of Fictionalist Suicide," takes what I have learned from thinking about moral fictionalism and applies it to another potential error-theoretic view: concerning the entities of folk psychology. It becomes quickly apparent that the psychological fictionalist faces some special problems that do not trouble other forms of the theory, and this paper attempts a fix. I should say that I'm not particularly inclined to doubt the existence of such things as beliefs and desires, though nor am I willing to declare that such doubt is misplaced. It is good to know, though, that even if one were to embrace such doubt, metaethical theories such as the error theory and fictionalism could still be identified, expressed, and advocated.

In preparing this collection I have resisted the temptation to mend any content which I now judge incorrect or at least think could be better expressed. (If I started down that road, where would I stop?) An exception is the addition of the first footnote to Essay 11, where I could not let my earlier self's claims go unchallenged. And on one

occasion I have corrected a reference to the optative mood to the cohortative mood. (What was I thinking?!) Apart from that, I have restricted myself to fixing typos, updating citations, and imposing a uniform spelling, punctuation, and formatting structure. I apologize for the occasional repetitions found in this collection; it is in my nature as a philosopher to go over the same ground frequently, and in this way gradually and incrementally make progress (or so the hope goes). The papers herein were produced while I was at the Australian National University, at the University of Sydney (which included a sabbatical near Périgueux), and then at Victoria University of Wellington. My thanks to the publishers of the various books and journals from which these papers are drawn.

Mitimiti, New Zealand
January 2015

PART I

Error Theory

1

Expressivism, Motivation Internalism, and Hume

David Hume is often taken to be a moral expressivist (Flew 1963; Ayer 1980: 84–5; Price 1988: 6; Snare 1991; Harman and Thomson 1996: 97). He is, moreover, often taken to have presented in the *Treatise* one of the strongest arguments for moral expressivism: the so-called *Motivation Argument*. As a metaethicist, I am interested in whether expressivism is true, and thus interested in whether the argument that people think they find in Hume is a sound one. Not being a Hume scholar (but merely a devoted fan), I am less interested in whether Hume really was an expressivist or whether he really did present an argument in its favor. Hume's metaethical views are very difficult to nail down, and by a careful selection of quotes one can present him as advocating expressivism, or cognitivist subjectivism, or moral skepticism, or a dispositional theory, or an ideal observer theory, or even utilitarianism. It is entirely possible that Hume's position is indeterminate when considered against these terms of modern moral philosophy; it is also entirely possible that he was hopelessly confused (much as it pains me to admit it). However, I doubt very much that Hume should be interpreted as an expressivist in any straightforward manner, and therefore I am doubtful that he should be interpreted as arguing in its favor. Most of this essay does not discuss Hume directly at all: I critically discuss the Motivation Argument and I advocate a certain positive metaethical view—one that mixes elements of traditional expressivism with elements of cognitivism. This position is neutral between moral realism and radical moral skepticism. I close by wondering—very briefly—whether Hume might have held such a view. Given my reservations about the determinacy of Hume's metaethical outlook, the case is not pressed with any vigor, but because it is an interpretation of Hume that has not, so far as I know, been articulated before, it may be of interest to note that it seems to be consistent with much of what he says—at least as much as any other precise interpretation.

Expressivism and Motivation Internalism

Let me start by clarifying terminology. *Noncognitivism* is the metaethical view according to which public moral judgments do not express beliefs (are not assertions) in spite of the fact that they are typically formed in the indicative mood. Thus defined,

noncognitivism is a view of what moral judgments *are not*—leaving open space for many different forms of noncognitivism claiming what moral judgments *are*. One positive form, *prescriptivism*, holds that moral judgments are really commands.[1] My focus in this essay is on another form, *expressivism*, which holds that moral judgments function to express desires, emotions, pro-/con-attitudes, or (in Simon Blackburn's words) "a stance, or conative state or pressure on choice and action" (1993: 168). I treat "expressivism" and "emotivism," as they appear in metaethical discussions, as synonyms.

Why might one be tempted by noncognitivism (and expressivism in particular)? First, noncognitivism sidesteps a number of thorny metaethical puzzles that face the cognitivist. The cognitivist thinks that when we make a public moral judgment, such as "That act of stealing was wrong," we are asserting that the act of stealing in question instantiates a certain property: wrongness. But queries arise: What kind of property is wrongness? How does it relate to the natural properties instantiated by the action? How do we have epistemic access to the property? How do we confirm whether something does or does not instantiate the property? The difficulty of answering such questions may lead one to reject the presupposition that prompted them: One might deny that in making a moral judgment we are engaging in the assignment of properties at all. Such a rejection, roughly speaking, is the noncognitivist proposal. Second, the noncognitivist might claim the advantage of more readily accounting for certain aspects of moral disagreement—for example, its vehemence and intractability (see Stevenson 1963, essays 1 and 2). The third traditional consideration in favor of noncognitivism is the subject of our attention: that noncognitivism does a better job than its rival of explaining the apparent motivational efficacy of moral judgment (see Smith 1994: chapters 1 and 2).

Those who advocate this third argument for noncognitivism often look to Hume for a precedent, finding solace especially in the following passage:

> Since morals, therefore, have an influence on the actions and affections, it follows, that they cannot be deriv'd from reason; and that because reason alone, as we have already prov'd, can never have any such influence. Morals excite passions, and produce or prevent actions. Reason of itself is utterly impotent in this particular. The rules of morality, therefore, are not conclusions of our reason. ([1740] 1978: 457)

It is not unusual to find Hume's premise that "morals excite passions" formulated as *motivation internalism*:

> *Simple Motivational Internalism (simple-MI)*:
> It is necessary and *a priori* that anyone who judges that she is morally required to ϕ will be (defeasibly) motivated to comply.

My object here is not to assess the truth of simple-MI, but to investigate its logical relation to metaethical expressivism. There are broadly two ways in which one might

[1] See, for example, Carnap (1935); Stevenson (1937).

use simple-MI in favor of expressivism. One might straightforwardly think that simple-MI entails expressivism, presumably with the supplement of some unobjectionable additional premise(s). Alternatively, one might think that simple-MI is a desideratum that any metaethical theory must strive to satisfy, and thus if expressivism were to entail simple-MI this would count very much in expressivism's favor. Let us examine both these implications in turn.

Does expressivism imply simple motivation internalism?

No, it does not. Expressivism, as I have stated it, is a thesis about what kind of mental states are expressed by moral judgments. It is vital to note that the notion of *expression* that is relevant here is non-causal. One can express a mental state while not having that state, and perhaps having never had that state. Consider a promise, which is a kind of speech act by which we express intentions or commitments. If I, in unexceptional circumstances, say to you "I promise to be at the party tonight," then I have expressed an intention to be at the party tonight. But my promise may nevertheless be insincere, in the sense that I have no intention of coming to the party, and have never had any intention to come. Since that intention appears nowhere in my mental repertoire, it *cannot be the cause* of my promise utterance. Now consider the following thesis:

> *Simple Promising Internalism*:
> It is necessary and *a priori* that anyone who promises to ϕ (thereby expressing the intention to ϕ) has the intention to ϕ.

It is clear that simple promising internalism does not follow from the thesis that promises express intentions. The phenomenon of insincerity is sufficient to demonstrate this. If, then, we construe metaethical expressivism as a thesis about what kind of speech act moral judgments are—which is natural if we read it as the denial that moral judgments are assertions, since *assertion* is a category of speech act (see Austin 1962; Searle 1969)—then the phenomenon of insincerity is entirely sufficient to show that simple-MI does not follow.

The reader might be wondering precisely what kind of relation is denoted by "expression" in this context if it is not a casual one. This is something I discuss later. One might also object that I have construed either expressivism or motivation internalism (or both) incorrectly. They are, I admit, both theses for which there is disagreement as to their correct formulation. I will consider variants in due course, but first let me consider the reverse implication with these simple formulations.

Does simple motivation internalism imply expressivism?

No, it does not. From the fact that there is a necessary (and *a priori*) connection between a kind of mental state and a kind of speech act, it does not follow that the speech act *expresses* that mental state. Let us consider promises again, and consider what criteria must be satisfied in order for X to succeed in making a promise (albeit

possibly an insincere one) to Y. The most complete answer to this question comes from John Searle (1969: 57–61), who painstakingly delineates the conditions that must obtain if *S* is to promise that *p* to *H* via uttering *T*. I shall not rehearse all Searle's items, but just focus on a couple. First: *H would prefer S's doing A to his not doing A*,[2] *and S believes H would prefer his doing A to his not doing A*. Second: *It is not obvious to both S and H that S will do A in the normal course of events*. Both these criteria make essential reference to the parties to the promise having certain *beliefs*. These connections are necessary and *a priori*: It is not possible that any person could succeed in making a promise to another person without their having these beliefs. Yet we would hardly say that the act of promising *functions to express the belief that the promisee would prefer that the promised action be performed to its not being performed*—rather, the mental state expressed by the promise is as we originally stated: an intention or commitment. This suffices to show that the occurrence of a type of speech act may entail that the speaker has a certain kind of mental state, though the speech act doesn't function to express that state.

A related confusion has cropped up in some quarters over what metaethical conclusions might be drawn from certain recent empirical results that show the important role that emotions play in moral judgment (see Greene et al. 2001; Greene and Haidt 2002; Moll et al. 2002; Haidt 2001). Although these scientists' conclusions are not uncontroversial, let us take them at their word when they assert that "recent evidence suggests that moral judgment is more a matter of emotion and affective intuition than deliberative reasoning" (Greene and Haidt 2002: 517). This conclusion is often referred to as "emotivism" (Haidt 2001: 816; Greene et al. 2001: 2107; Greene et al. 2004: 397). Anthropologist Daniel Fessler claims that "emotivist perspectives on moral reasoning hold that emotional reactions precede propositional reasoning" (Fessler et al. 2003: 31). Let us be a little bolder, and interpret this kind of emotivism as the claim that all moral judgments are *caused by* emotional arousal.

Clearly, this use of "emotivism" among empirical scientists is very different from the metaethicist's usage, for whom it is usually a synonym of "expressivism."[3] The terms "emotivism" and "expressivism" *in the metaethical tradition* do not denote a thesis about the causal origins of moral judgment; they denote (as we have seen) a thesis about what kind of mental state is expressed by public moral judgments. It might be best if we distinguish "psychological emotivism" (the kind advocated by Jon Haidt, for example) from "metaethical emotivism" (advocated by A. J. Ayer and Simon Blackburn, for example).[4] The crucial point is to note the logical independence of the

[2] "*A*" denotes a future act that proposition *p* predicates of *S*.

[3] See Joyce (2008) for further discussion of this disparity between psychologists' and metaethicists' use of "emotivism."

[4] Páll Árdal (1966) once distinguished "emotionism" from "emotivism," in a way that maps closely to the contrast between what I am calling "psychological emotivism" and "metaethical emotivism" (though perhaps it is even closer to the distinction that I will make shortly, between the mentalistic construal of expressivism and the metaethical construal of expressivism). I quite like Árdal's terminology, but it never took off. Prinz (2007) has recently reintroduced the word "emotionism" for a somewhat different thesis.

two: Even if the evidence were to demonstrate that every single moral judgment is caused by emotional arousal (that is, demonstrate that psychological emotivism is true), this wouldn't imply anything about *the function of* moral language.

Consider, for example, any kind of metaethical theory according to which moral utterances are veiled reports about one's own mental states. According to this kind of cognitivist subjectivism, "X is morally wrong" means "I feel disapproval toward X."[5] The latter is something that may be asserted, yet in the typical case it will have been prompted by emotional arousal in the speaker. Such a theory would be consistent with psychological emotivism but inconsistent with metaethical emotivism. On at least one occasion, Hume sounds like he endorses some such view: "[W]hen you pronounce any action or character to be vicious, you mean nothing, but that from the constitution of your nature you have a feeling or sentiment of blame from the contemplation of it" ([1740] 1978: 469). I doubt that this particular simplistic subjectivist interpretation of Hume ultimately withstands scrutiny, but I have little doubt that Hume should nevertheless be interpreted as some kind of psychological emotivist: He clearly and emphatically thinks that moral judgments have their origin not in the faculty of reason but in sentiment.[6] My point is that his advocacy of psychological emotivism does not commit him to the metaethical variety; there is very little evidence that he advocated, or even had much awareness of, metaethical emotivism/expressivism/noncognitivism. Many of his emotivist-sounding moments (for example, "Morality . . . is more properly felt than judg'd of" ([1740] 1978: 470)) may be smoothly interpreted as advocating psychological emotivism rather than metaethical. Indeed, in the earlier-quoted passage from the *Treatise* from which the Motivation Argument is drawn, what we in fact have, I would claim, is an argument for psychological emotivism, not metaethical emotivism.

In sum, there is clearly a significant difference between motivation internalism and psychological emotivism. One asserts a necessary connection between moral judgment and conative states, whereas the other asserts a causal connection between the two. But advocates of either must be wary of the same potential pitfall: of assuming that their thesis implies or provides support for the thesis that metaethicists have called "emotivism" or "expressivism." I have argued that expressivism does not imply motivation internalism, and nor does motivation internalism imply expressivism. The latter

[5] Charles Stevenson (1937; 1963) held a view of this sort, though he maintained that in addition the moral judgment includes a command.

[6] John Bricke writes of the subjectivist-sounding *Treatise* passage (3.1.1.23): "The autobiographical rendering of evaluative sentences being so utterly implausible, it is fortunate that there is no reason whatever to think that Hume here means by 'meaning' what, when concerned with language, we now mean" (1996: 162). (See also Ayer 1980: 84.) Nicholas Sturgeon (2008: 514) interprets Hume as a subjectivist, but not the speaker-oriented kind mentioned here. After quoting the subjectivist-sounding passage, Sturgeon notes that Hume subsequently "modifies this view to make the truth of one's ascription of virtue or vice depend, not on one's actual feelings, but on the feelings one would have under the right conditions, whether or not one now is (or even could be) in those conditions." I am not sure whether Sturgeon's Hume ultimately counts as a psychological emotivist. This is my own fault, since I have characterized psychological emotivism only as carefully as is necessary to reveal its logical independence from metaethical emotivism—but have left it intentionally indeterminate in several respects.

denial raises serious problems for an argument that some people have found in Hume: the Motivation Argument. Rather than claiming that Hume argued poorly, however, I urge the conclusion that it is a mistake to foist this argument for expressivism on him in the first place.[7]

Some will object to all of the preceding on the grounds that I have misconstrued motivation internalism or misconstrued expressivism (or misconstrued both). So it is to variations on these theses that I now turn.

Variants of Expressivism and Motivation Internalism

A prominent variation on motivation internalism is the following, due (*inter alios*) to Michael Smith (1994):

> *Normative Motivation Internalism*:
> It is necessary and *a priori* that anyone who judges that she is morally required to ϕ will be (defeasibly) motivated to comply, or she is irrational.

I mention this variant simply because it is well known, but in fact there is little to say about it here. The additional clause on the end, though important in other philosophical contexts, does not affect any of the arguments that I have already deployed to show the logical independence of expressivism and simple motivation internalism. This is less obviously so of the following variant:

> *Sincerity Motivation Internalism (sincerity-MI)*:
> It is necessary and *a priori* that anyone who sincerely judges that she is morally required to ϕ will be (defeasibly) motivated to comply.

(One may choose to add the suffix "…or she is irrational"; it does not matter to anything that follows.) The thesis is not always worded in just this way, but the term "sincerely" is often included in statements of motivation internalism (see Hare 1999: chapter 8; Timmons 1999: 53; Svavarsdóttir 2006: 186; Shafer-Landau 2005: 142).

Sincerity-MI does not imply expressivism. Proof: Sincerity-MI is implied by simple-MI, therefore if sincerity-MI were to imply expressivism, then so too would simple-MI imply expressivism. But we have already seen that simple-MI does not imply expressivism, therefore nor does sincerity-MI. Thus, construing MI as sincerity-MI provides no succor for the (alleged) Humean Motivation Argument for expressivism.

But does expressivism imply sincerity-MI? One might be tempted to think so. Suppose for the sake of argument that expressivism is true: that when one (in ordinary circumstances) utters "Stealing is morally wrong" (say), one thereby expresses some

[7] Rachel Cohen concurs that Hume's Motivation Argument "is irrelevant to non-cognitivism" (1997: 261); however, her positive interpretation differs from mine: She maintains that the argument concerns the nature of moral properties.

conative (that is, motivation-implicating) state. One might think that from this follows something about what it is for such a judgment to be sincere. One might think that any conation-expressing utterance of "Stealing is morally wrong" is sincere if and only if the speaker actually has that conative state at the time of utterance. This would be an instance of a tempting general principle of speech-act sincerity, which I shall name after its advocate, John Searle (1969):

Searle-Sincerity:
S's utterance U (at time *t*) is sincere iff U expresses mental state M, and S has M (at *t*).

With Searle-Sincerity as an additional premise, it does appear that expressivism implies sincerity-MI. The problem is that Searle-Sincerity, plausible as it may appear at a glance, is false.

On an earlier occasion (Joyce 2002) I offered some counterexamples to Searle-Sincerity. I imagined someone saying "Thanks!" as he left a dinner party in a distracted and hurried way, and claimed that though we might admit that at the moment of thanking he was feeling no gratitude whatsoever, nevertheless we would not ordinarily call his utterance "insincere." A second counterexample along the same lines concerned an act of passing moral judgment. Michael Ridge (2006) criticizes these counterexamples on the grounds that they fail to take into account the fact that although the speaker may not have gratitude (say) as an *occurrent* emotion, he nonetheless may count as having that mental state (at the time of utterance) *dispositionally*. I harbor misgivings about dispositional mental states (especially emotional ones), but let us not pause to consider them now, for the main point is that Ridge nevertheless agrees with me that Searle-Sincerity is inadequate, and supplies counterexamples of his own that revolve around self-delusion. A person may believe himself to have mental state M when in fact he does not. If there is a speech act that expresses M, and the person performs that speech act, then it seems natural to say (Ridge argues) that the speech act is sincere, even though the speaker lacks the mental state that it expresses.

Ridge presents an alternative general thesis of speech-act sincerity:

Ridge-Sincerity:
S's utterance U (at time *t*) is sincere iff S believes that U expresses mental state M, and S believes that she has M (at *t*).[8]

I am concerned that Ridge's version of speech-act sincerity is also problematic, in that it presupposes that ordinary speakers have beliefs about a kind of expression-relation holding between utterances and mental states—but this, I suspect, is far too *recherché* a belief to require of ordinary speakers in order that they may be granted speech-act sincerity (even if we allow that the belief may be implicit, dispositional, and non-conscious). We have already seen the confusions that may entrap the

[8] For the sake of brevity I have stripped Ridge's thesis of a few details that don't matter on this occasion. For the full account see Ridge (2006: 501).

unwary concerning causal versus conventional notions of *expression*; and if even analytic philosophers stumble over this, what hope should we have that ordinary speakers' beliefs are in order? Even children of a tender age have the capacity to make assertions, ask questions, bark commands (and so forth)—and to do so in a sincere manner—yet surely they have no beliefs about what kind of mental states various utterances *express*. I once had an argument with a well-known philosopher (who shall remain nameless) who declared that metaethicists have no idea what they are talking about when they wonder about what mental states moral judgments *express*. He claimed that it was as if metaethicists had just assumed the existence of some mysterious relation holding between moral utterances and mental states, then given it a name—"expression" (though it may as well have been "floog")—and have then expended endless energy arguing in circles about this baffling relation. This philosopher presumably (or at least conceivably) did *not* believe that any of his utterances *expressed* any mental states; he was sufficiently skeptical of the whole notion that he just withheld assent to such thoughts. Do we want to claim that this philosopher, despite himself, "implicitly" had such beliefs? I wouldn't wish to maintain this; it seems a rejoinder of desperation. I might allow that his speech acts, despite his beliefs, *did* express various mental states, but I see little plausibility in the claim that, despite himself, he *believed* this fact. And yet, for all this, I am quite certain that this person was capable of making sincere assertions, sincere promises, sincere apologies, and so forth.

Here I am not going to argue for an alternative general theory of speech-act sincerity, since I question the assumption that a general account is forthcoming or even particularly desirable. Perhaps what must be added to a promise to ensure its sincerity differs from what must be added to an assertion to ensure its sincerity, while both differ from congratulations, apologies, entreaties, thankings, and so on. If there is anything that unites these cases, in my opinion, it will revolve around the fact that insincere speech acts are ones by which the speaker attempts knowingly to mislead his/her audience—and such audience-directed intentions are not mentioned, nor entailed, by either Searle-Sincerity or Ridge-Sincerity. But I shall not develop this thought on this occasion, for the point that matters to our present purposes does not require it. For our present purposes I can even embrace Ridge-Sincerity. The point is that expressivism promises to imply sincerity-MI only with Searle-Sincerity as a bridging premise, but Searle-Sincerity has been refuted, and there is no reason to assume that whatever general thesis of speech-act sincerity replaces it (if there even is one) will also act as a bridge from one thesis to the other.

Of course, if we have any account whatsoever of what a sincere moral judgment consists of, it will follow trivially that *some* kind of vaguely motivation-internalism-ish thesis will be implied by expressivism. Consider the generalized argumentative format:

Premise 1: Moral judgments express conative state C (regarding the subject of the judgment).

Premise 2: Sincere moral judgments have quality Q (vis-à-vis the mental state expressed).

Therefore: If S sincerely judges that she is morally required to ϕ, then her judgment has quality Q vis-à-vis C (regarding her ϕing).

I call the conclusion "motivation-internalism-ish" on the grounds that it asserts a relation between moral judgment and a kind of motivational state (C). (And if the two premises are necessary and *a priori* then the relation asserted in the conclusion will also be so.) But in fact this conclusion is derived trivially, and has no useful role to play in the metaethical dialectic. Recall that the expressivist hope has been that some version of motivation internalism might have independent attractions, thus boosting the case for expressivism either by implying expressivism or by being revealed to be a desideratum that expressivism satisfies better than its rivals. But the kind of trivial MI-ish conclusion just mentioned can play neither role.

Another obvious problem with any version of motivation internalism that restricts itself to *sincere* moral judgments is that by encompassing only a proper subset of moral judgments, it fails to tell us anything about moral judgment *simpliciter*. By comparison, I could tell you something true about all moral judgments made on a Saturday—say, that they are "weekend moral judgments"—but this obviously would tell us nothing about what we are interested in as metaethicists: namely, what a moral judgment *is*, what its necessary features are, and so on. Similarly, the fact that sincerity-MI reveals a connection between sincere moral judgments and motivational states doesn't imply any necessary connection between motivation and moral judgments *simpliciter*, since the motivational aspect may be smuggled in within the concept of *sincerity*. (If I am correct that the most promising account of speech-act sincerity will make reference to the speaker's intentions not to deceive, then sincerity will automatically bring motivation along for free, since these kinds of intentions are motivation-engaging states.) Any version of internalism restricted to sincere moral judgments is compatible with the falsity of expressivism concerning moral judgments *simpliciter*.

I have lately been discussing variations on the thesis of motivation internalism, but it may also be objected that I have misconstrued the thesis of expressivism. One might, in particular, complain about my characterization of expressivism as a metaethical theory about *speech acts*; one might instead insist that expressivism is a theory about *mental states*: not about what kind of mental state moral judgments *express*, but about what kind of mental state moral judgments *are*. On such a view the applicability of the sincere/insincere distinction retreats and, indeed, the whole troublesome expression relation conveniently evaporates.

This mentalistic construal of expressivism is unconventional. If we go back to the roots of noncognitivism in the early twentieth century, we see pretty clearly that what is under discussion is the nature of moral *language*. In their influential 1923 book *The Meaning of Meaning*, C. K. Ogden and I. A. Richards speak of a use of the word "good" which is "purely emotive," and "[w]hen so used the word stands for nothing whatsoever,

and has no symbolic function" (1923: 125). A. J. Ayer's noncognitivism was motivated by the question of how moral utterances might be meaningful *statements* ([1936] 1971).[9] Rudolf Carnap's noncognitivism was presented as the claim that a "value statement" is not "really an assertive proposition," but is, rather, "a command in misleading grammatical form" (1935: 24–5). Charles Stevenson spoke of ethical judgments as having "quasi-imperative force" which may be "intensified by your tone of voice" (1937: 19). At no point in these classic works in the emergence of noncognitivism is it hinted that "moral judgment" might be used primarily to denote a mental state.

But perhaps that is all a misguided historical idiosyncrasy, and perhaps we would do better now to treat moral judgments as a species of mental state. In this case, expressivism will be the theory that moral judgments are not beliefs, but rather some kind of conative state (to be specified). One obvious problem with such a decision is that it opens the possibility that moral judgments (qua mental states) might be conative, while moral judgments (qua linguistic entities) might be assertoric. This is more or less the same possibility as was noted earlier, when I pointed out that metaethical cognitivism (interpreted in the orthodox manner) is compatible with either simple-MI or psychological emotivism. Of course, this observation doesn't count as evidence against mentalistic expressivism, but rather indicates how confusing this way of characterizing theories might become.

In the present context, the important thing to note is that construing expressivism mentalistically would nullify the possibility of any argumentatively interesting relation holding between expressivism and motivation internalism. If expressivism is the theory that moral judgments *are* episodes of conative state C (where "C" denotes something that is by stipulation necessarily motivation-engaging), then expressivism is essentially equivalent to the thesis of motivation internalism, which states that moral judgments necessarily engage motivation. The connections appear to be so trivial that arguing for either thesis by means of first establishing the other ceases to be a feasible dialectical strategy. This has particular relevance to the so-called Motivation Argument that is drawn from Hume's *Treatise*. From the premise that "morals excite passions" (that is, putatively, that moral judgments necessarily engage motivations) one can certainly derive mentalistic expressivism—but only trivially: The conclusion essentially *is* the premise.[10, 11]

[9] Ayer thought that all meaningful statements must be either analytic or empirically verifiable. Given that moral utterances appear to be neither, Ayer was forced to claim that they are not meaningful statements. But rather than concluding that moral judgments are meaningless, Ayer's preferred conclusion is that they are not *statements*, but are, rather, ways of evincing ones emotions and issuing commands.

[10] It is possible that one might construe MI and mentalistic expressivism such that they have a different modal and/or epistemological status. MI, recall, is presented as a necessary and *a priori* thesis; perhaps mentalistic expressivism need not be. It is difficult, however, to see how this would create the possibility of a viable argumentative strategy from one to the other. From "It is actually the case that X" we cannot conclude "It is necessarily the case that X." The reverse implication does hold, of course, but then the question is on what grounds we could establish the necessity claim as the antecedent. If we had any such grounds, then the consequent would hardly be in doubt.

[11] Frank Snare interprets the argument from motivation as aiming to establish mentalistic expressivism (he calls it "emotivism"). He concludes that the influence of the argument is due to philosophers having

One can read the preceding as presenting a dilemma to any expressivist tempted to employ the Humean Motivation Argument. Either expressivism is construed mentalistically (as a theory of what kind of mental states moral judgments *are*), in which case the argument is valid but question begging, or it is construed linguistically (as a theory of what kind of mental states moral judgments *express*), in which case the argument is unsound. I have made clear my preference for construing expressivism in the latter fashion (thus impaling the advocate of the Motivation Argument on the second horn), and, in the course of discussion, I made much of the fact that the relevant expression relation must not be understood in a causal manner. I should like now to say more in a positive vein about how that expression relation ought to be understood. This discussion takes us well away from Hume—which is hardly surprising given my contention that thinking of Hume in the guise of a modern expressivist is a serious distortion—but I will close with some brief thoughts applying what we have learned to Hume.

Expressing Mental States

When seeking explication of the sense in which types of speech act express types of mental states, it is useful to start with Moore's Paradox. G. E. Moore (1942: 543[12]) noted the oddity of someone's claiming:

(1) I went to the pictures last Tuesday. But I don't believe that I did.

It's called a "paradox" because although it is not a logical contradiction (for it is perfectly possible that I went to the pictures last Tuesday while I don't believe that I did), to *state* the whole is to void the speech act of the first part, leaving the listener confused as to what should be assumed about the speaker's attitude toward his having been to the pictures last Tuesday. It makes (to quote J. L. Austin) "a peculiar kind of nonsense" (1961: 235). Moore presents the paradox using the category of assertion, but it seems we can find exactly the same phenomenon with other species of speech act:

(2) I apologize for having lied to you. But I have absolutely no regret about having lied to you.
(3) Thank you for the present. But I have no gratitude toward you for giving me the present.

One can imagine someone's publicly uttering the first component of any of (1), (2), and (3) while uttering the second component *sotto voce*—in which case the public speech act (assertion, apology, and thanking, respectively) would simply be insincere. There is

"been so completely convinced of the conclusion that they did not realize that the conclusion itself provides much of the reason for believing the premises" (Snare 1975: 9).

[12] I am not sure whether Moore mentioned the paradox in print on any earlier occasion. Wittgenstein reports Moore having mentioned it in a lecture, probably from before the First World War. (Norman Malcolm mentions that Wittgenstein opined that this was the only work of Moore's that had ever "greatly impressed him" (Malcolm 1958: 66).)

nothing very noteworthy about that. What is strange is when the second component is uttered *out loud* along with the first—since a speech act that wears its insincerity on its sleeve is apt to cause confusion.[13] It is, to quote Austin again, "a statement that fails to get by" (1963: 28). Note that I am not merely saying that (1)–(3) "sound odd to my ear," but am trying to locate a particular kind of oddity, which promises to help us understand the relevant notion of *expressing* a mental state. David Copp introduces a useful phrase to denote this kind of expression relation: "Moore-expression" (correspondence 2003; see his 2001: 10). It is illuminating, I think, to consider these matters when reflecting on pejorative language, such as racial slurs.

(4) Aaron's a kike. But I have no contempt toward him or people of his ethnicity or religion.[14]

I am inclined to think that (4) is a manifestation of the same phenomenon as the rest, though Copp would disagree. Although he would concur that calling someone a "kike" certainly in some sense expresses contempt, and that someone who uttered (4) should expect to be challenged to explain, he judges that it doesn't *Moore-express* contempt, since the implication can be cancelled. (Copp's example concerns calling a canine a "cur" as opposed to a "mongrel dog"; I am assuming that it is a fair analogy.) One could add further explanatory comments to (4) so as to assuage the audience's confusion in the following manner: "Aaron is a kike; but I have no contempt toward Aaron or people of his ethnicity or religion; it's just that for the moment I've forgotten the usual non-derogatory term for such persons." Thus, Copp thinks, the contempt expressed by "kike" may be cancelled. Perhaps he is right, though I confess to feeling uneasy about having to rely on an example of someone forgetting a word or learning a language in order to illustrate the context of cancellation. Is there *anything* that is non-cancellable if we admit such odd contexts?

Consider the case of the Reverend William Spooner, who famously concluded one of his sermons with the addendum: "In the sermon I have just preached, whenever I said 'Aristotle' I meant to say 'St. Paul'." Suppose the sermon had included the assertion of the sentence "Aristotle was born in Tarsus" (when in fact it was St Paul who was born there), and, when queried on this by a bewildered student, Spooner had responded (coming momentarily to his senses, but oblivious of what he had just been saying) "No, of course Aristotle wasn't born in Tarsus; he was born in Stagira!" Thus we have what appears to be a flat contradiction asserted—"Aristotle was born in Tarsus

[13] Compare the phenomenon of sarcasm: If Fred looks out at the pelting English rain and sighs "another glorious day in paradise" in tones of sarcasm clear for all to hear, then he has not asserted that it is a glorious day in paradise. But sarcasm is such an entrenched convention that his audience will not be confused—they will know that he did not assert it, and did not intend to assert it, and Fred will know that they know. Thus sarcasm is not a case of insincerity "worn on the sleeve," since it is not a case of insincerity at all.

[14] A lot depends on tone of voice here. It does with (1)–(3) as well, since uttering any of these in tones of sarcasm or in a clearly joking way completely alters the speech acts involved. Let me stipulate that we are to imagine these uttered in serious tones. (Insert the word "stinking" before "kike" if it helps to reinforce this point.) I apologize to anyone who finds the term offensive even in the context of *mentioning the word as an example of an extremely offensive word*. On another occasion of making this argument, I employed the less provocative term "kraut"—but the very choice of a word whose offensiveness is less deeply ingrained undermined the cogency of the argument.

and it is not the case that Aristotle was born in Tarsus" (the first conjunct uttered as part of the confused sermon, and the second conjunct, being an answer to a question, not part of the sermon)—but one made intelligible by the addendum admitting a linguistic confusion. This possibility hardly undermines the fact that "Aristotle was born in Tarsus" is semantically implied by "Aristotle was born in Tarsus," and that this implication is as non-cancellable as they come! Recall that Paul Grice confessed to the unreliability of the cancellability test when speakers are talking "in a loose or relaxed way" (Grice 1989: 44), and I suspect that imagining a person who has an imperfect grasp of the language, or who is forced to employ unconventional words because she cannot remember the correct ones, renders the test null.

We will return to the question of the status of (4) shortly, but for now apply these thoughts to a moral utterance:

(5) Hitler was evil. But I do not believe that he was evil.

This seems a straightforward instance of Moore's Paradox, which strongly suggests that an ordinary freestanding utterance of "Hitler was evil" is an assertion. If an utterance of the form "But I don't have mental state M" can nullify the speech act that the preceding comment would otherwise perform, then it is natural to assume (*ceteris paribus*) that the preceding comment functions to express M.

So does noncognitivism stand refuted? Not by any means, for it seems that we can observe the same phenomenon, or at least a very similar one, if in (5) we substitute for "belief" something conative. Which conative state we choose, and how we opt to describe it, may make a significant difference here, and there are various options. Copp (following Allan Gibbard) prefers to see the expressed conative state as *acceptance of, or subscription to, a moral standard* (Copp 2001: 30). According to this stipulative use of "subscription," if a person actually does take this attitude, "she is in a state of mind that, if effective, constrains and guides her planning so that she is motivated to some degree [to comply]" (Copp 2001: 30). Since no proposal more plausible suggests itself, let us use the following wording:

(6) Hitler was evil. But I subscribe to no normative standard that condemns him or his actions.

I'm cautiously tempted to treat (6) the same as the others, to hold that all six of the numbered sentence pairs so far considered are Moore-paradoxical, meaning that in each the relation between the speech act of the first sentence and the mental state mentioned in the second sentence is one of Moore-expression. If this is true of (5) and (6), then the conclusion to be drawn is that an ordinary freestanding utterance of "Hitler was evil" expresses *both a belief and a conative attitude*. This would speak in support of a metaethical view that mixes aspects of traditional noncognitivism with components of traditional cognitivism—a view I favor.[15]

[15] Such a mixed view has been maintained by C. L. Stevenson, R. M. Hare, and P. H. Nowell-Smith, among others. If one defines noncognitivism simply as the denial of cognitivism, then the two theories are,

Copp, however, though agreeing that (1), (2), (3), and (5) reveal instances of Moore-expression, thinks that (6) should go with (4)—as *not* Moore-paradoxical—for the reason mentioned earlier: that the strangeness of (6) may be assuaged via additional commentary, and thus the customary relation between the utterance and the mental state in question may be cancelled. He would find nothing unintelligible about an amoralist who says "Hitler was evil; but I subscribe to no normative standard that condemns him or his actions. I just don't go in for morality; I believe in it all right, but I think it's a manipulative cultural invention that is best avoided." I have two responses to Copp's claim here. The first is to deny that the amoralist's declaration is intelligible *in the relevant way*, and thus maintain grounds for holding that (6) is Moore-paradoxical. The second is to accept that (6) is not Moore-paradoxical, but explore whether there is another kind of *expression* involved—a kind that, while not being Moore-expression, is nevertheless robust enough to underwrite the noncognitivist element of moral judgments. Either avenue is sufficiently promising that the prospects for the mixed metaethical view to which I have just adverted seem encouraging. Discussion of the former response calls for a section to itself, after which I will turn to the latter response.

Amoralist Cancellation

There is, of course, a sense in which all of the numbered sentence pairs (1)–(6) are intelligible. They are grammatical and non-contradictory sentences. But so too is any Moore-paradoxical sentence pair (and many nonsensical sentences besides). Yet I maintain that they are unintelligible in the sense that someone hearing any such pair (someone who has not been primed in some special way, that is) would be unsure about what speech act has been performed by the first component.[16] Considering (6): Copp is correct that any perplexity that may be aroused may also be assuaged with further commentary—the kind of "but-I-don't-go-in-for-morality" comment that was mentioned. We are familiar with this kind of an amoralist; after all, didn't Plato's Thrasymachus proclaim something along similar lines?[17]

of course, contradictories. Similarly, if one defines cognitivism as the theory that moral judgments express *only* beliefs, then any "mixed theory" will be excluded. However, if we think of expressivism as a positive proposal (about what moral judgments *do* express), and we drop the "only" clause in both expressivism and cognitivism, then moral judgments may be *two* things: They may be assertions *and* ways of expressing conative attitudes.

[16] Compare John MacFarlane (2005: 334), who writes: "Imagine someone saying: 'I concede that what I asserted wasn't true, but I stand by what I said anyway.' We would have a very difficult time taking such a person seriously as an asserter. If she continued to manifest this kind of indifference to established truth, we would stop regarding the noises coming out of her mouth as assertions."

[17] Note how "amoralism" is a term of art here. In the vernacular, "amoralist" often denotes someone who rejects morality altogether, who doesn't believe in it at all. In recent philosophical debates, by contrast, it denotes someone who makes genuine moral judgments but lacks any motivation to comply (and the topic of the debate is whether the amoralist is even a possibility). I am using the term in a third way: to denote someone who tries to cancel the motivation-implicating aspect of a moral judgment. The debate here is not whether such agents exist, but whether they succeed in making moral judgments.

Yet care needs to be taken in our treatment of these pieces of additional commentary that render intelligible preceding speech that would otherwise be confusing. There's a sense in which *any* verbal nonsense can be rendered intelligible with the addition of "…and what I just uttered was a great example of verbal nonsense." The crucial matter is whether the additional intelligibility-instating commentary leaves intact the apparent speech act performed by the first sentence. For example, suppose that A utters after a dinner party "Well, *the plates* were nice"; we might presume that it is being conversationally implicated that the food was unpleasant. But this implicatum may be smoothly cancelled, if A were to add: "Of course, I don't mean to imply that the food was unpleasant; it was nice too." Whatever wisps of strangeness might remain hanging over this pair of utterances will be dissipated if A explains: "I was just so taken by the plates that for a moment I wasn't thinking about the food." The important thing to notice is that at the end of all this explaining we are content that A's first sentence was indeed what we thought it was: an assertion that the plates were nice. By comparison, if B says "The cat is on the mat, but I don't believe it," then goes on to add "and that's a good example of confusing language," then although listeners may be comfortable with the total exchange, they will not know whether B has asserted that the cat is on the mat. The crucial question, then, is not whether the Thrasymachian amoralist who renders intelligible (6) by explaining that he "just doesn't go in for morality" is someone we can make sense of (I concede that he is); the question is whether after the intelligibility-reinstating amoralist explanation the audience is confident that the speaker has made a genuine moral judgment. And about this I think there is substantial doubt.

Before discussing the amoralist further, it will be useful to remind ourselves of how fluid and scrappy linguistic conventions can be. Let me draw attention to four general points.

First, a solid linguistic convention may be quite easily overridden by another. There is little doubt that the term "slut" functions as a pejorative in English. Yet by introducing the overarching convention of *joking*—which may be achieved in a second by a shift in tone or a twitch of an eyebrow—one might in a playful manner say to a close female friend "Oh, you're such a slut" with all offensiveness nullified. Yet even in these circumstances "slut" continues to be a contempt-expressing term, for that, after all, is what makes the comment funny.

Second, many terms that function to express attitudes as well as beliefs will also have purely belief-expressing uses as well. The word "queer" in the sense of *unusual and peculiar* remains neutral even if "queer" in reference to homosexuality can be used as a term of derision. The word "bastard" for a long while could be used descriptively to mean *illegitimate offspring*, even when it could also be employed as a term of insult. Similarly, most, if not all, of the terms centrally associated with moral judgment also can be used non-morally. Possibly the most fundamental term of moral appraisal is "ought." The "ought" that appears in "Mary morally ought to refrain from stealing" may express the speaker's subscription to norms that condemn stealing, but nobody is claiming this of the weather forecaster's utterance of "It ought to rain tomorrow."

Similarly, to acknowledge that someone is "a good assassin" is not to express any kind of endorsing attitude, whereas to claim that someone is "a good person" is.

A third point to note about such conventions is that they can change very quickly. The terms "idiot," "moron," and "cretin" were once respectable scientific labels; the term that largely replaced them—"mentally retarded"—is at present the subject of controversy. There will be transitional times when one should not claim with confidence either that there is or is not a linguistic convention according to which the term expresses an attitude.

A fourth point is that linguistic subgroups can create linguistic sub-conventions. Within certain gay circles, referring to one's gay friends as "queer" may be neutral, though it may be highly insulting for an outsider to select that term. Much the same could be said about the use of the word "nigger" among some African American subgroups. In some circles, to call something "bad" (in a certain tone of voice, perhaps) is a way of praising it, and in surfing lingo, to call a wave "wicked" is to express admiration for its qualities. Such conventions may be sequentially embraced and overridden within a single conversation—or, indeed, within a single sentence: "Which of you bastards called this bastard a bastard?"[18]

What I hope these observations call attention to is the fact that although we might be able to *imagine* someone intelligibly advocating the amoralist line, it doesn't show that there is not actually an entrenched convention according to which the use of moral terms expresses subscription to a norm. What we need to ask ourselves is whether any such imaginative act involves us thinking of aberrant subgroups, or people speaking in a joking, playful manner, or the speaker using something like a sarcastic or ironic tone of voice, or using a moral term in a non-moral manner, or introducing a new convention by example, or a world with slightly different linguistic conventions than we actually do have, or so on. When Milton's Satan says "Evil, be thou my good" the natural reading is that Satan is doing something tricky with language. A careful analysis of his comment would take too long here; it's enough to note that although we know exactly what Satan is trying to communicate, we also recognize that the surface construction is paradoxical (and this, of course, is what gives the line its poetic power). Thus, that Satan's comment should be intelligible doesn't reveal that there is not actually a linguistic convention according to which to call something "evil" is to express one's subscription to a standard that condemns it. It is exactly this convention that Milton has exploited in a clever and mischievous way.

I believe the same thing can be said quite generally of the amoralist's apparent cancellation of the conative component of a moral judgment: "Hitler was evil; but I subscribe to no normative standard that condemns him or his actions. I just don't go in for morality; I believe in it all right, but I think it's a manipulative cultural invention that is

[18] This sentence was reportedly uttered by the Australian cricket captain during the "bodyline series" of 1932–33. The English captain came to the Australian dressing room to complain about one of his players having been called a "bastard" during play. Bill Woodfull, the Australian captain, turned to his team and uttered the memorable line. (I owe this example to the late David Lewis (correspondence 2000).)

best avoided." Although a person would be *intelligible* if she said this, it seems to me that we would be left in serious doubt as to whether she has really *judged Hitler to be evil*. And the reason for this indecision, I think, is precisely that the careful explanation that the speaker offers of her position reveals that a degree of stipulative usage is being introduced. She is explicitly suspending a convention that is in place in regular language. But the fact that one can do this, and do it with ease, hardly shows that there is not actually such a convention, any more than the fact that I can say "For the next few minutes I will use the word 'cat' to stand for dogs" (or "Cats, be thou my dogs") reveals that there is some doubt concerning whether in English "cat" denotes cats.[19]

Frege-Expression

I have just argued in favor of treating the relation between moral judgment and certain conative states as an instance of Moore-expression. The issue of where the line should be drawn between Moore-paradoxical and non-Moore-paradoxical utterances is difficult to settle, since the rules for how we should restrict the contexts in which cancellation may or may not be possible are undecided. But suppose that Copp is correct that (4) and (6) are not examples of Moore-paradox. Does it follow that there is no expressivist component to pejorative slurs and moral judgments? Copp doesn't think so, and nor do I.

Drawing inspiration from Frege's views on "coloring," Copp claims that pejorative terms and moral terms *Frege-express* mental states. Frege wrote that two words might have the very same sense and reference, and yet one might lend the utterance a "coloring" (*Färbung*) that the other does not, such that choosing to use one word rather than the other (for example, "kike" rather than "Jewish person") might be "unsuitable, as if a song with a sad subject were to be sung in a lively fashion" ([1892] 1997: 167). Frege's own example involves the word "cur": He writes that "whilst the word 'dog' is neutral as between having pleasant or unpleasant associations, the word 'cur' certainly has unpleasant rather than pleasant associations and puts us in mind of a dog with a

[19] There is a well-known interpretation of the amoralist from R. M. Hare (1952: 124–6, 167ff), according to which the amoralist's statement is not literally a moral judgment at all, but rather is best read as having quotation marks round the term "morally ought." My view is not unlike this, though it is important to bear in mind that Hare's amoralist utters something like "For me to steal would be wrong" while *having* no motivation to refrain from stealing, while my amoralist says, "For me to steal would be wrong, though I subscribe to no normative framework that condemns stealing." Given the careful verbal qualification that the latter offers, it seems to me quite plausible that something rather like quotation marks are being imposed. Copp objects that Hare's view fails to accommodate the possibility of moralists and amoralists entering into moral debate (Copp 2001: 13). If the amoralist says "Liberalism is a great evil" and the moralist responds "No, liberalism is morally defensible," but in fact the former statement is equivalent to something like "It is considered hereabouts that liberalism is a great evil," then there is no real disagreement. But I do not find this the reductio that Copp seems to think it to be. Perhaps any intuition we have that there can be genuine moral debate with amoralists just stems from the fact that the way they speak (making the quotation marks tacit) is apt to encourage us to forget that they are not really making moral judgments at all. If I were to hear someone claim that liberalism is a great evil then I would want to protest; but if I were then to discover that this person had earlier asserted "Evil, be thou my good," then I should become quite confused as to what she thought about liberalism, and thus not at all confident that I should disagree.

somewhat unkempt appearance" ([1897] 1997: 240–1). When such coloring becomes an entrenched custom in the linguistic community—as is the case with words like "kike" or "slut"—then we can, according to Copp, consider the relation between the utterance and the proposition that the speaker has the attitude in question (for example, contempt) to be a variety of conventional implicature. Note, though, that Copp thinks that this expressiveness is cancellable, and in this he diverges from Grice, for whom, apparently, non-cancellability is a feature by which conventional implicatures are to be distinguished from conversational implicatures. Although cancellable, these "colorings" are a type of conventional rather than conversational implicature (for Copp) because they are detachable[20] and because to employ a colored term while lacking the attitude in question would be a *misuse* of the term. At the risk of sounding evasive, I prefer to sidestep the Gricean framework, if only because it strikes me as sufficiently unclear and contested that one only courts controversy in trying to apply it to new domains. I have already expressed my misgivings about the cancellability of colorings, but even if I agreed with Copp on this point, it seems to me imprudent to employ the term "conventional implicature"—a term of art partially defined by reference to *non-cancellability*—to categorize the phenomenon.[21]

Nevertheless, I am strongly inclined to agree with Copp (and Frege) on the general point that colorings are, or at least can be, a matter of *linguistic convention*. This may be a vague claim, but it is good enough for my present purposes. The contemptuous attitude of someone who uses "kike" rather than "Jewish person" is not merely an expectation that interlocutors will have formed on the basis of past observation. In teaching the word "kike" to a novice language-user, it would be intolerably negligent to refrain from mentioning the term's evaluative baggage. Indeed, someone who didn't know that "kike" was a contempt-expressing term could legitimately be said not to understand the term properly at all, even if able competently to apply it to all and only Jewish people. Any such ignorant person would not require any tutoring concerning what it takes to be Jewish, but is in need of *linguistic* instruction.

Although the comparison threatens to be misunderstood if taken too far, I agree with Copp that moral language is in important respects like pejorative language. More precisely: The way that moral judgments express conative attitudes is very similar to, if not the same as, the way that pejorative terms express attitudes. The manner in which (6) fails to get by may or may not be precisely the same as the manner in which (5) fails to get by, but it is close enough as to make no difference to the general conclusion that I am trying to reach: that moral judgments express, as a matter of entrenched linguistic convention, both beliefs *and* conative attitudes. The traditional debate between the

[20] For Grice, the implicatum *p* is detachable from an utterance iff there are ways of saying the same thing that do not implicate *p* (see Grice 1989: 39). What Copp has in mind in saying that colorings are detachable is that instead of "Aaron is a kike" one could say "Aaron is Jewish," and the latter, though saying the same thing (having the same sense and reference?), lacks the implicatum that the speaker has contempt.

[21] Copp himself admits that "nothing [in my argument] turns on whether coloring is an example of conventional implicature or simply a phenomenon that is similar to conventional implicature" (2001: 23).

cognitivist and the noncognitivist has not taken into account such nuances as the distinction between Moore-expression and Frege-expression—indeed, has been scandalously casual about what it means to say that such-and-such judgments *express* mental state so-and-so—and thus, were we ultimately to conclude that moral judgments Moore-express beliefs and Frege-express attitudes, this could not be construed as a victory for either party. My main point is that a modicum of reflection on the issue reveals the traditional metaethical debate between the cognitivist and the noncognitivist to rest on a false dichotomy.

Hume: Expressivist, Cognitivist, *and* Skeptic?

Earlier I claimed that there is very little evidence that Hume advocated expressivism. This was not entirely true; the point I was trying to press is that the places where Hume has traditionally been read as promoting expressivism (or noncognitivism more generally) should not be construed that way. For the real hints of expressivism in Hume, one must look to where he discusses evaluative *language*. As a preliminary, we should remind ourselves that Hume did have at his disposal a remarkably forward-looking account of how indicative language can be used in non-assertoric ways. His sophisticated discussion of promising foreshadows J. L. Austin's:

> [T]here is *a certain form of words*…by which we bind ourselves to the performance of any action. This form of words constitutes what we call a *promise*…When a man says *he promises any thing*, he in effect expresses a *resolution* of performing it. ([1740] 1978: 522)

Regarding evaluative language, Hume is (occasionally) clear that there are entrenched conventions that associate conative states with certain words:

> Every tongue possesses one set of words which are taken in a good sense, and another in the opposite. ([1751] 1998: 6)

> [W]hen [someone] bestows on any man the epithets of *vicious* or *odious* or *depraved*, he…expresses sentiments, in which, he expects, all his audience are to concur with him. ([1751] 1998: 75)

> [T]here are certain terms in every language which import blame, and others praise; and all men who use the same tongue must agree in their application of them…This great unanimity is usually ascribed to the influence of plain reason, which…maintains similar sentiments in all men…But we must also allow, that some part of the seeming harmony in morals may be accounted for from the very nature of language. The word *virtue*, with its equivalent in every tongue, implies praise, as that of *vice* does blame; and no one, without the most obvious and grossest impropriety, could affix reproach to a term, which in general acceptation is understood in a good sense; or bestow applause, where the idiom requires disapprobation. ([1757] 1996: 134–5)

It is in such passages—few and far between as they are—that we find Hume the expressivist. But these comments are presented in a way that makes clear that Hume considers them as peripheral to any of his central arguments; he evidently does not, on these

occasions, take himself to be putting forward any weighty and controversial metaethical thesis in need of argumentative support. In other words, to the extent that Hume is an expressivist, it is not something he thinks worthy of making a song and dance about; he hardly notices that he is taking (what we would now classify as) a metaethical stance. More importantly, note that these expressivist musings have no obvious role to play in the moral thesis that Hume is really obsessed with: that morals are the product of a sentimental faculty rather than a rational faculty (a thesis I earlier called "psychological emotivism"). In particular, these passages are very far indeed—both logically and textually—from the much-touted *Motivation Argument* for expressivism.

What I hope to have shown in the previous two sections is that even if Hume does have expressivist leanings, this does not exclude his also robustly endorsing a cognitivist metaethical view. Moral cognitivism comes in both realist and skeptic flavors, and each of these possibilities is compatible with expressivism. Copp, for example, articulates and advocates a position he calls "realist-expressivism." He thinks that the truth conditions for the belief expressed by a moral judgment such as "Cursing is wrong" concern cursing being prohibited by a "relevantly justified or authoritative moral standard or norm" (2001: 27). What it takes for a moral standard to be appropriately "authoritative" is, in the first instance, left open. Copp's own view is a *society-centered theory*: that a standard is authoritative just in case "its currency in the social code of the relevant society would best contribute to the society's ability to meet its needs—including its needs for physical continuity, internal harmony and cooperative interaction, and peaceful and cooperative relations with its neighbors" (2001: 28). Since, we may assume, such justification is sometimes forthcoming, in Copp's view moral judgments will turn out sometimes to be true. He combines this realism with the thesis that moral language Frege-expresses conative states; hence: realist-expressivism.[22]

I should like to draw attention to another branch of the tree of metaethics: *error-theoretic expressivism*. There are different ways that one might argue for this position, but it is convenient to use Copp's view as a point of departure. Suppose he is correct that the cognitive element of a moral claim refers to something like the "relevantly justified or authoritative moral standard or norm." One may, nevertheless, think that Copp's preferred explication of *justification* is too relativistic or too anthropocentric to capture the kind of practical authority we demand of a moral theory. For the kind of familiar reasons outlined by John Mackie, for example, one might think that inherent in moral discourse is a commitment to a kind of institution-transcendent practical categoricity that is in fact not satisfied by anything in the world (Mackie 1977; see also Joyce 2001).[23] Thus one might agree with Copp concerning how to understand

[22] I actually harbor some reservations that Copp's view quite deserves the label "realism," but the fact that it is a version of moral cognitivist "success theory" is enough to underwrite the distinction I am highlighting. For my views on how to characterize *moral (anti)realism*, see Joyce (2007a).

[23] Indeed, Mackie's general definition of "good" is not a million miles away from the cognitivist element of morality articulated by Copp. Mackie defines "good" as "such as to satisfy requirements (etc.) of the kind in question" (1977: 55–6). With many non-moral uses of "good" Mackie thinks the predicate is satisfied.

the expressivist element of moral discourse, and also agree with his views concerning the truth conditions (*broadly* construed) of moral judgments, while holding that these truth conditions are never satisfied. Hence: error-theoretic expressivism.

Might Hume be a realist-expressivist or an error-theoretic expressivist? I would not want to press either case with any confidence, but I will nevertheless close with a brief exploration of Hume's commitment to the moral error theory. Note that I do not claim for a moment that Hume thought of himself as an error theorist but just expressed himself poorly, nor even that he would have embraced the view had it been articulated to him. But there are certainly threads in Hume's moral philosophy that lean in that direction.[24]

In looking for evidence *against* this interpretation, one might bring forth any of a number of Hume's critical comments aimed at moral skepticism. But it should be remembered that the kind of skeptic whom Hume has in mind is invariably the *Pyrrhonic* skeptic: someone who thinks that we cannot know whether claims of a certain kind are true or false and therefore ought to withhold passing judgment on the matter. The error theorist, by contrast, is no Pyrrhonic skeptic, but (in classical terms) should be classified as a negative dogmatist (or nihilist). Similarly, many of Hume's comments apparently targeting moral nihilists (for example, "those who have denied the reality of moral distinctions" ([1751] 1998: 3)) on more careful examination seem to be admonishing those who would pretend *indifference*, who would claim not to *care* whether a person was honest or a thief. But this is also something that a moral error theorist may distance himself from. The moral error theorist may be as opposed to tax fraud, as sickened by pedophilia, as horrified by genocide, as anyone else. Error-theoretic moral skepticism implies nothing about how *tolerant* its advocates will be.

Hume is opposed not merely to Pyrrhonic ataraxic indifference toward morality, but is averse to any suggestion that philosophizing should lead us to *give up* the practice of making moral judgments. I can discern no hint of moral eliminativism in his writings. The widespread assumption that eliminativism is the natural consequence of a moral error theory may have something to do with a reluctance to press the error-theoretic interpretation of Hume. This assumption, however, is flawed. There may be pragmatic reasons for maintaining moral thought and moral language even once moral skepticism has been embraced (see Joyce 2001; Kalderon 2005a). Or it may be that the human mind is simply *unable* to give up these practices, even when philosophical considerations have led one to see the flaws. One can find allusions to the former in Hume's writings, and the latter is something of a recurring theme. In his essay "The Sceptic," he notes "that famous doctrine" that colors exist not in nature but

But in *moral* contexts, he thinks, the pertinent requirements are those that are "simply there, in the nature of things, without being the requirements of any person or body of persons, even God" (1977: 59). It is Mackie's conviction that there are no such "intrinsic requirements" that leads to his moral skepticism.

[24] Mackie (1980) argues along similar lines to myself for the error-theoretic interpretation of Hume. David Gauthier (1992) also toys with this interpretation of Hume's account of the artificial virtues, though does not firmly endorse it.

only in the eye, then poses the rhetorical question: "[If this were so,] would dyers or painters ever be less regarded or esteemed?" He goes on to ask "why should a like discovery in moral philosophy make any alteration?" ([1742] 1996: 354). The closing section of Book 1 of the *Treatise* waxes lyrical about how simply *living* life will drive all skeptical musings from one's mind:

> I dine, I play a game of back-gammon, I converse, and am merry with my friends; and when after three or four hour's amusement, I wou'd return to these speculations, they appear so cold, and strain'd, and ridiculous, that I cannot find it in my heart to enter into them any farther. ([1740] 1978: 269)

In his abstract for the *Treatise* (in which he writes of himself in the third person), Hume sums up this view:

> Our author insists upon several other sceptical topics; and on the whole concludes, that we assent to our faculties, and employ our reason only because we cannot help it. Philosophy would render us entirely *Pyrrhonian* were not nature too strong for it. ([1740] 1978: 657)

Of anyone who would profess indifference to moral distinctions, Hume counsels that the best response is "to leave him to himself," trusting that "it is probable he will, at last, of himself, from mere weariness, come over to the side of common sense and reason" ([1751] 1998: 3). In other words, even if one came to espouse a moral error theory, "nature herself" would eventually drive that philosophical allegiance from one's mind. And Hume evidently thinks that this would be no bad thing. Of those "honest gentlemen of England"—who "being always employ'd in domestic affairs, or amusing themselves in common recreations, have carried their thoughts very little beyond those objects"—Hume tells us that "they do well to keep themselves in their present situation" ([1740] 1978: 272). Yet even if these down-to-earth folk are acknowledged to be well off, Hume is not attempting to dissuade anyone from engaging in philosophical speculations: He simply thinks that one either will or will not, according to temperament and mood, and that to the extent that one will it is likely that one's efforts will be temporary (that is, until someone calls out "Anyone for backgammon?").

Yet none of this is at odds with the possible *truth* of an error theory. Regarding causal relations and the continued existence of external objects, Hume is explicit that experience "leads us into errors" ([1740] 1978: 267) (for he thinks that the two beliefs are jointly affirmed contraries), but even here—where we seem to have Hume clearly endorsing some kind of error theory—he thinks it remains an open question the extent to which he should "torture my brain with subtilities" ([1740] 1978: 270), the extent to which he should "yield to these illusions" ([1740] 1978: 267). Nature may ineluctably reassert herself against the awareness that one has fallen into error, but the errors are no less errors for that.

Might things stand similarly for morality, in Hume's eyes? He certainly never claims outright that morality does not exist. He is more likely to say something along the lines of "[m]oral properties exist not in bodies but merely in the senses." It is worth noting

that such claims remain consistent with a moral error theory. Telling someone "The pink elephants exist only in your mind" is in fact a way of saying that the pink elephants do not exist at all. At one point in the *Treatise* Hume declares that sounds and smells "really exist no where" ([1740] 1978: 167), and it is reasonable to think that he will say the same of color and causation (the section in question concerns "necessary connexion"). This bald claim of non-existence comes immediately after he has spoken of the mind's "great propensity to spread itself on external objects"—a thesis that I will here refer to as "projectivism." The other well-known projectivist passage from Hume is in *An Enquiry Concerning the Principles of Morals*, where he claims that "taste" (as opposed to reason) "has a productive faculty, and gilding and staining all natural objects with the colours, borrowed from internal sentiment, raises in a manner a new creation" ([1751] 1998: 89). Here he mentions "beauty and deformity, vice and virtue" as the products of sentimental projection. The correlation of these two projectivist passages suggests that what goes for one (that is, that sounds and smells "really exist no where") should go for the other (that is, that virtue and vice "really exist no where"). That moral qualities should receive the same treatment as color, sound, smell, heat, and cold is reaffirmed elsewhere in the *Treatise* ([1740] 1978: 469) and in "The Sceptic."

Hume is, moreover, explicit that the folk do indeed think of colors, smells, sounds, and heat as objective qualities of objects. In a letter to Hugh Blair of 1762 he is dismissive of the idea that the folk might not be objectivists; evidently, "the Vulgar" (as Hume refers to them) are taken in by their own projectivist tendencies. "Philosophy scarce ever advances a greater Paradox in the Eyes of the People, than when it affirms that Snow is neither cold nor white: Fire hot nor red" ([1762] 1986: 416). On the assumption that what is said here will carry over for moral qualities as well, then the folk are generally fooled by their moral projectivist tendencies: They are unaware that their moral judgments are the product of sentiments being projected onto the world; they both experience the world as morally "colored" *and believe it to be*. But if someone *believes* something to be the case, then it is natural to assume that her utterances on the matter will be *assertions*. Thus, I claim, Hume's moral projectivism is a form of psychological emotivism that leads naturally to (but I would not go so far as to say *implies*) moral cognitivism—though a cognitivism that remains compatible with an expressivist component.[25]

But are the assertions in question *true*? If snow is not white—something that Hume seemingly endorses—and someone asserts the sentence "Snow is white," then the very natural conclusion to draw is that she has simply asserted something false. Similarly, if Fred's character does not have the quality of *virtuousness*, and someone asserts the sentence "Fred is virtuous," then the natural conclusion is that she has asserted something false. One way to avoid this error-theoretic interpretation of Hume's metaethical commitments is to problematize the analogy that allows us to draw conclusions about his

[25] In saying this I am flying in the face of recent tradition, which tends to lump projectivism together with noncognitivism. I have argued elsewhere that this is at best optional and at worst a mistake. See Joyce (2006a: chapter 4), and Joyce (2009a). In the latter, I delineate different species of projectivism.

implicit moral views from what he somewhat more explicitly claims about sensory modalities. I have nothing to say here on that score, except to reaffirm that Hume draws the analogy sufficiently frequently that it is safe to assume that he thinks there are illuminating similarities among these topics. Another way is to leave the analogy intact but to deny the error-theoretic construal of both analogs. Perhaps when Hume denies that snow is white what he means is something like "Snow—*considered in itself, restricting ourselves to the consideration of only its intrinsic qualities*—is not white." But perhaps he also thinks that it is not obligatory to understand the truth conditions of "Snow is white" in this manner. If we allow the possibility of whiteness being some relational, subject-implicating property, then perhaps Hume will consider the sentence true, after all. The problem is that he has adamantly affirmed that the general folk think of the whiteness of snow in the former objectivist fashion. The crux of the issue, then, is whether the weight of dominant folk belief on the matter is sufficient to determine how the word "white" must be understood (or, if you prefer, how the identity conditions of the concept *whiteness* must be construed). Understood one way it leads to an error theory; understood in another way it leads to a success theory.

Hume had no resources for addressing this question, and nor have we. This leads to an impasse in establishing whether Hume's occasional expressivist tendencies are mixed with a cognitivism that is committed to success or a cognitivism that is committed to skepticism. The contemporary debate between the moral error theorist and the moral success theorist seems locked up at exactly the same point: A problematic (or "queer") quality of morality is brought forth, regarding which some will argue that this quality is an *essential* aspect of the moral conceptual framework (such that any normative system stripped of this problematic element would no longer deserve the name "morality"), whereas others will argue that extirpation of the flawed element would amount merely to a benign revision and demystification of morality. Lacking an accepted methodology for deciding such disputes, the modern metaethical debate is at a disappointing stalemate.[26] It is also possible that on occasions there is no fact of the matter about whether a given discourse (for example, morality) is committed to some putatively queer property, leading to the conclusion that the dispute between the moral error theorist and her many detractors may in fact be fundamentally undecidable.[27] If this is correct, then it is tempting to suppose that any interpretation of Hume that leaves him sitting on the fence over this matter is a charitable one.[28]

[26] I discuss this impasse in Joyce (2006a: chapter 6, 2007a, 2007b).

[27] David Lewis writes: "Strictly speaking, Mackie is right: genuine values would have to meet an impossible condition, so it is an error to think there are any. Loosely speaking, the name may go to a claimant that deserves it imperfectly . . . What to make of the situation is mainly a matter of temperament" (1989: 136–7).

[28] Early portions of this chapter closely follow Joyce (2002). A youthful version of this essay once went by the name of "Noncognitivism, motivation, and assertion," and it was helped along by feedback from David Lewis and Simon Kirchin. I thank David Copp for the very useful discussion.

2

Morality, Schmorality

In his contribution to this volume,* Paul Bloomfield analyzes and attempts to answer the question "Why is it bad to be bad?" I too will use this question as my point of departure; in particular I want to approach the matter from the perspective of a moral error theorist. This discussion will preface one of the principal topics of this essay: the relationship between morality and self-interest. Again, my main goal is to clarify what the moral error theorist might say on this subject. Against this background, the final portion of this essay will be a discussion of moral fictionalism, defending it from some objections.

Bloomfield is correct to claim that the best way of removing the appearance of tautology or poor formation from the question "Why is it bad to be bad?" is to gloss it as elliptical for something along the lines of "Why does being morally bad have a deleterious effect on my self-interest?" The two "bad"s are intensionally non-identical: One (I will assume) refers to a non-moral notion of prudential badness (whatever is, all things considered, harmful to one's welfare[1]), while the other refers to a kind of ostensibly distinct *moral* badness. Although both notions have enough intuitive meat to them for discussion to proceed, neither is unproblematic. (I will return to these problems later.) On this interpretation, the question "Is it bad to be bad?" can be seen as an inquiry concerning whether two intensionally nonequivalent concepts are such that in fact (or even, perhaps, necessarily) the extension of one includes the extension of the other.[2] To ask the question "*Why* is it bad to be bad?" is to presuppose that this *is* the case, and to inquire in virtue of what this is so.

The moral error theorist thinks (1) that the predicate "…is morally bad" is a logical predicate (in contrast to the semantic noncognitivist, who thinks that it is a predicate only in a grammatical sense), (2) that sentences of the form "ϕ is morally bad" are

* P. Bloomfield (ed.), *Morality and Self-Interest* (Oxford University Press, 2007). All references to "this volume" in this essay are to Bloomfield's collection, in which this essay originally appeared.

[1] Contra W. D. Falk (this volume), in this essay I am not using the term "prudence" to denote to a policy of risk avoidance, but rather am identifying prudence with acting in whatever way advances one's interests, all things considered. (And the relevant notion of "interests" I am leaving unspecified.) I am happy also to use the term "expediency" as a synonym for the same. Despite Falk's insistence that "expediency" must implicate some notion of *convenience*, my dictionary tells me that it also means simply "self-serving."

[2] I take it that nobody will claim that the two concepts are coextensive—that every act of imprudence is a moral wrong. My having a cup of coffee before going to bed may be prudentially foolish, but surely doesn't count as even a mild moral crime.

generally uttered with assertoric force (in contrast to the pragmatic noncognitivist, who thinks that such sentences are used to perform some other linguistic function), and (3) that the predicate "... is morally bad" has an empty extension (in contrast to, for example, the moral realist, who thinks that the property of moral badness is instantiated).[3] The third contention is the most controversial, and there are various reasons that might lead one to endorse it; it is not my intention in this essay to attempt to make any of these reasons compelling. Perhaps the error theorist thinks that for something to be morally bad would imply or presuppose that human actions enjoy a kind of unrestricted *autonomy*, while thinking that in fact the universe supplies no such autonomy (Haji 1998, 2003). Perhaps she thinks that for something to be morally bad would imply or presuppose a kind of inescapable, authoritative imperative against pursuing that thing, while thinking that in fact the universe supplies no such imperatives (Mackie 1977; Joyce 2001). Perhaps she thinks that for something to be morally bad would imply or presuppose that human moral attitudes manifest a kind of uniformity, while thinking that in fact attitudes do not converge (Burgess [1978] 2010). Perhaps she thinks that there exists no phenomenon whose explanation requires that the property of moral badness be instantiated, while thinking that explanatory redundancy is good ground for disbelief (Hinckfuss 1987). Perhaps she thinks that tracing the history of the concept *moral badness* back to its origins reveals a basis in supernatural forces and magical bonds—a defective metaphysical framework outside which the concept makes no sense (Hägerström 1953). Perhaps she thinks all of these things and more besides.[4] The details are not important here; the point is that the error theorist accuses morality of being fatally flawed, such that any value system with the flawed element(s) extirpated simply wouldn't deserve the name "morality." The only detail that need be noted here about the moral error-theoretic position is that it is usually restricted to the moral realm. Of course, in principle one could endorse a radical *global* error theory, in which case one would by implication be an error theorist about morality (along with modality, colors, other minds, cats and dogs, and so on), or one could be an error theorist about all normative phenomena, which, again, would include an error theory for morality. But typically the moral error theorist thinks that there is something *especially* problematic about morality, and does not harbor the same doubts about normativity in general. The moral error theorist usually allows that we can still deliberate about how to act, she thinks that we can still make sense of actions harming or advancing our own welfare (and others' welfare), and thus she thinks that we can continue to make sense of prudential "ought"s.[5] She allows that prudential badness is instantiated but insists that

[3] The options mentioned in this sentence are not intended to be exhaustive.

[4] For the sake of brevity I will talk as if the error theorist thinks there is only one thing problematic about morality. But of course an error theorist may be impressed by a number of considerations against morality. Perhaps morality has a lot of little or medium-sized problems—none of which by itself would ground an error theory, but all of which together constitute A Big Problem.

[5] In this essay I assume that prudence naturally takes the form of a normative system, that it involves "ought" claims, reasons for action, and so on. In fact, one could deny this. All that is minimally necessary for believing in prudence is to accept that individuals can be harmed. Thus even if one thought that all

moral badness is not. Thus, on the assumption that the question "Is it bad to be bad?" amounts to an inquiry about the truth value of a universal conditional ("Is it the case that: For any *x*, if *x* is morally bad, then *x* is prudentially bad?"), the moral error theorist will think that the answer to the question is vacuously "Yes" because the conditional has a false antecedent irrespective of how the variable is instantiated.[6] (Note that she will also, for the same reason, answer "Is it good to be bad?" in the positive.) Thus, she will object to the presupposition behind the question "Why is it bad to be bad?" In this respect the question is, for her, not unlike "Why is it bad to annoy a witch?" Her answer is "But you *can't* annoy a witch—there aren't any!"

But there is another way of understanding the elliptical element of the question that allows the possibility of the error theorist giving a substantive and interesting "Yes." If she treats the reference to moral badness as denoting the extension *that it is widely assumed to have*, the extension is not empty at all (though see below). After all, the error theorist is well aware that there is a broad range of actions—both types and tokens—that are widely thought to be morally bad: breaking promises, stealing, unprovoked violence, Hitler's Final Solution, gluttony, sloth, envy, and so on and so forth.[7] She can understand the question "Is it bad to be bad?" as "Will performing *these* actions [gesturing to those actions that are widely considered to be morally bad] have a deleterious effect on the interests of the perpetrator?" For token actions that have already been performed—for example, Ernie's lying to Bert last week—the question must be either "Did this action have a deleterious effect on Ernie's interests?" or "Would performing an action of the same type have a deleterious effect on the interests of the perpetrator?" These are all questions that the error theorist might answer positively, thus allowing that the question "*Why* is it bad for me to pursue such things?" must have an answer.

(By comparison, suppose an anthropologist were studying a culture in which certain persons are considered to be witches. The anthropologist might recognize that it's a good thing—good for his research, that is—if he stays on friendly terms with these persons, even though he doesn't believe that they possess the supernatural powers necessary for actually being witches at all. He might say "It is good to be friendly to those persons that are hereabouts considered to be witches," but there would be nothing impermissible, or, in general, misleading, if he were, for convenience, to express this elliptically as "It is good to be friendly to the witches hereabouts.")

One problem with this interpretation is that there may be significant disagreement among the people "hereabouts" as to what counts as morally bad, such that even the predicate "…is widely assumed to be morally bad" threatens to turn up empty. After

"ought" claims are false—even all non-moral ones—one could still uphold that "…is prudentially bad" has a non-empty extension.

[6] There are some complications here concerning (A) whether the domain of the variable is restricted to actual entities, and (B) whether the error theorist holds that moral predicates are necessarily empty or just actually empty. Addressing these complications is unnecessary.

[7] Although for the sake of simplicity I tend to speak just of actions being morally bad, I don't mean to exclude morally bad character traits, states of affairs, intentions, policies, properties, objects, and so forth.

all, moral discourse, it is often observed, is characterized by a high degree of intractable disagreement. Perhaps, though, there are at least *some* things for which there is sufficient concurrence that we can speak of "what is widely assumed to be bad" (strangling babies?), and perhaps the error theorist confines her question merely to these actions. Or perhaps the error theorist just passes the buck to her interlocutor, and says: "Tell me what things you consider to be morally bad, and I will tell you whether (and, if so, in virtue of what) their pursuit is imprudent."

Another feature of the question to which attention should be drawn is the fact that it may receive different answers for different people, or for the same person at different times, or for the same person (or counterparts, if you prefer) at different possible worlds. Perhaps it will frustrate Ernie's interests to lie, but it will not frustrate Bert's interests to lie. Perhaps it will frustrate Ernie's interests to lie today, but he'll be okay if he waits till next Friday. Or perhaps it will frustrate both Ernie's and Bert's interests to lie, but it will do so for very different reasons: For example, Ernie would have to live with crippling guilt, whereas Bert would be sent to bed without any dinner. Or perhaps as a matter of fact everyone has a prudential reason to avoid badness (and perhaps they all have the same reason), but there are possible circumstances where the pursuit of the bad would become prudentially good (for at least some persons).

It has been a longstanding aspiration of a certain school of moral philosophy—upon whose roll appears the name "Bloomfield, P."—that all such contingent messiness could be swept aside by the provision of a universal, permanent, monolithic and (perhaps) necessary positive answer to the question. Bloomfield's solution is that all bad human agents undermine their self-respect and thus frustrate their own interests. As far as go the principal theses of this paper, Bloomfield may be entirely right. But I happen to doubt that he is, and I find my sense of courtesy to the good editor of this volume prevailed over by an intellectual urge to join the fray; hence I cannot forego making a couple of critical comments.

First, it should be noted that at best his argument shows that there is something self-damaging about a certain kind of radical, ubiquitous, all-encompassing, self-conscious attitude toward what is (widely assumed to be) bad: pleonexia. But whether there even *are* any such awful characters around is a moot question. The agents who perform those actions widely thought of as bad—breaking promises, stealing, and so on—are rarely inclined to appeal to Thrasymachian or Machiavellian iconoclasm to attempt to justify themselves. Most everyday wrongdoers,[8] I submit, believe that what they are doing *isn't* really bad (and that if others disagree it's because they're not properly acquainted with the details of the case). Wrongdoing is born of negligence as often as it is born of arrogance. Many wrongdoers castigate themselves for their actions, and even perform them regretfully. Wrongdoers are not always selfishly

[8] Here I am using the term "wrongdoers" in a purely descriptive manner: to pick out those people who are widely considered to be wrongdoers. Not wanting to beg the question against the error theorist, I should really keep the term in scare quotes throughout, but I refrain from doing so for stylistic reasons.

motivated by gewgaws: Consider a mafia hitman acting out of obligation and loyalty (perhaps even love) for the paterfamilias. Few wrongdoers fail to distinguish between the out-group (a domain of potential victims) and the in-group (a domain of friends, family, loved ones, those with whom one has binding obligations, and so on). In short, the pleonectic may be a fascinating philosophical case study, but he hardly represents the typical or paradigmatic instance of badness.

Bloomfield seems to think that the pleonectic represents the toughest case, and thus that if even Thrasymachus and his ilk can be shown to be harming themselves, then surely all those more mundane wrongdoers must proportionally follow suit. But this expectation is, in my opinion, ill-founded. Although the pleonectic does in some sense occupy an extreme wing of villainy, it doesn't follow that any injury he does himself must by implication be suffered to a lesser degree by less radical wrongdoers. Consider, for example, the claim that the pleonectic lives a life without "true love" and lacking "real friendship." We can all accept that any human who chooses such an existence is very probably damaging himself. But what about a lesser wrongdoer who, say, is creative with the truth when filing his taxes, or is needlessly curt to a taxi driver? It might be claimed that this person has harmed himself to a lesser degree by missing an opportunity for *some* true love and *some* real friends (that is, the love and friendship of the victims, perhaps). But there is surely nothing wrong with this kind of loss per se, for *everyone*—even the thoroughly virtuous—must eschew *some* potential friendships. (I don't recall that Mother Teresa ever sent me a Christmas card.) I see no grounds for assuming that a mundane wrongdoer cannot enjoy the full complement of genuine friends, or that the occasional bit of everyday misconduct (directed at non-friends) must, to some small degree, undermine those friendships. Consider instead the claim that certain pleonectics must be guilty of psychological compartmentalization. We can all agree that extreme compartmentalization of one's thoughts and desires is a harmful state. But what about *a little* compartmentalization? There is presumably nothing wrong with "a bit" of compartmentalization, since, again, it is an attribute that *every* human exhibits; it's the nature of human psychology. It might be complained that in this context the term "compartmentalization" is intended to denote only the pernicious, pathological variety. But then we are free to deny that the mild transgressor must manifest *any* such attribute, and any insistence that he does so simply begs the question.

If I am correct that the harm that the pleonectic (allegedly) does himself derives from aspects of the very extremism of his attitude, then there are no grounds for thinking that a lesser degree of the same kind of self-harm is in store for the everyday moral transgressor. And thus we have not been shown how moral badness per se is self-injurious, but rather only how a proper subset of moral badness is bad—and a very small (and perhaps actually empty) proper subset of moral badness at that.

The second critical comment I will make against Bloomfield's argument is that it at best shows that there is some kind of *fault* with the pleonectic, but it is not clear how this fault translates into an injury. The pleonectic, according to Bloomfield, has but a

simulacrum of self-respect; what she takes to be self-respect is "faulty in its foundations." Because the pleonectic accords others no respect, she cannot coherently respect herself, for to do so would be based on the (allegedly) impermissible distinction that "I deserve more because I am me." Although the pleonectic may be quite convinced that he does have self-respect, he is, in fact, self-deceived. To grant Bloomfield this case (something that I am, in fact, very far from doing) would be to acknowledge that a milestone in philosophy has been achieved. An argument demonstrating the irrationality of wrongdoers is something that Simon Blackburn has described as the "holy grail of moral philosophy" (1984: 222). Nevertheless, even if Bloomfield's argument delivered the grail into our hands, this would not achieve the goal he set himself, for it is simply not clear how being irrational or self-deceived entails doing oneself harm. "I am special because I am me" may be a misguided or irrational thought (though even this I am highly doubtful of), but why *self-harmful*?

If someone is habitually irrational in all her deliberations then it is not unreasonable to suppose that this will land her in various kinds of trouble; and it is not hard to see that self-deception will often be self-injurious. But to show that irrationality and self-deception are on very many occasions harmful is insufficient to establish that there is anything harmful about these phenomena per se. This is especially evident when the charge of irrationality/self-deception is so unobvious that it takes a philosopher to establish it—against a background of more than 2,000 years of like endeavors meeting with a body of staunch academic opposition. When the accusation concerns so inconspicuous and subtle a phenomenon, then one must suspend any assumption that the typical harms that issue from canonical and obvious irrationality/self-deception must also issue from the inconspicuous instances. In other words, if Bloomfield were to succeed in demonstrating that every moral wrongdoer is to some extent self-deceived, then he would have shown us that the domain of self-deception is very different than it is widely assumed to be, and thus any previous assumptions about the general harmfulness of self-deception (based, as they are, on a different class of prototypes) would stand in need of re-examination.

Just as space allowed Bloomfield to make his case but briefly, so too I will not attempt to respond to his final "five things that could be said to Thrasymachus" in any detail. My main suspicion is that they are indeed things that could be said *to Thrasymachus* (that is, to the pleonectic), but have considerably less force against a more everyday wrongdoer. That someone who cheats slightly on his taxes, or is needlessly discourteous to the taxi driver, is suffering from schizophrenia, that he must endure the anxieties of dissimulation, that he is missing the "the joy of seeing things as they actually are"—that he is leaning toward any of these wretched states *even slightly*—is, at best, an optimistic claim in need of empirical support. Of course there is a kind of satisfaction that comes from a job done with moral integrity; but there is also a satisfaction that comes from getting away with something. Of the people who have experienced both, of course there are some who prefer the first kind of pleasure; but there are also, I'll wager, those who prefer the second kind. (Many of us are not

insensitive to both kinds.) The latter people may very well be self-deceived—it is not my intention here to deny it—but what needs to be asked is whether they are *harming themselves*. To appeal to a "joy" that comes from having true beliefs may sound appealing—especially to a philosopher—but I don't think it stands up to scrutiny. Do true beliefs *always* bring this joy? I don't recall the last time I felt even a hint of ecstasy when contemplating that 1 + 1 = 2. Perhaps Bloomfield means to restrict his comment just to a certain domain of epistemic success: a joy that comes from having true beliefs about our own value in comparison to that of other humans. Again, Bloomfield's opponent need not deny the very possibility of such a joy, nor even deny that it might be quite widespread. All she need deny is the universal claim that Bloomfield's argument requires if it is to succeed: that such joy is available to *anyone* in *any* circumstance, and that it can never be outweighed by a countervailing joy that flows from gaining benefits (and not necessarily mere gewgaws) secured through an act of moral transgression.

It is not my intention to criticize Bloomfield's argument beyond these gestures, because the main point to which I want to draw attention is that as far as the moral error theorist is concerned Bloomfield could be 100 percent correct. Chances are, what the moral error theorist is likely to say in response to the question "Is it bad to be morally bad?" (understood as outlined above) is "Sometimes it is, sometimes it isn't." But were she instead to answer "Yes: always, for everyone, necessarily"—and then go on to justify this answer by appeal to Bloomfield's argument—she would in no sense undermine her commitment to a moral error theory. Embracing a moral error theory rationally eliminates from one's serious practical deliberations certain kinds of justification: One can no longer, for example, refrain from doing something because one believes that it is morally forbidden. But it implies nothing about what actions one should actually perform (or refrain from performing). Contrary to popular belief, the moral error theorist is not a scheming villain, acting pleasantly solely in order to avoid punishment or to lull her victims into complacency. (As Richard Garner puts it: "The amoralist need not be an immoral, heartless, selfish jerk who denies the obvious" (1994: 279).) The moral error theorist may have as much compassion, love, and generosity as anyone else; she will just not believe these characteristics, or their attendant actions, to be *morally* desirable.[9] Nor does the embrace of a moral error theory obviously exclude any particular *non-moral* forms of justification from figuring in one's deliberations. The moral error theorist may be motivated largely by compassion, or by self-interest, or by a sense of loyalty to her friends and family, or (more likely) a mixture of these things (and others besides) depending on

[9] We mustn't be distracted by the fact that such emotions as love and generosity are often called "*moral* emotions." If they warrant this label it is in virtue of the fact that they are considered morally praiseworthy, but it is clear that one can have these emotions without making any moral judgment. The moral error theorist does not have her position undermined if *others* choose to judge her character, actions, and emotions in moral terms.

the situation.[10] There is simply no reason to assume that having such a (non-moral) basis to one's deliberations is going to end up prescribing sneaky nastiness. On the contrary, for most people, in most ordinary situations, it is fair to assume that a proper sensitivity to such non-moral considerations is likely to favor acting in accordance with (what most people think of as) moral requirements. So the moral error theorist is as willing and able as anyone else to endorse claims such as "I ought not break promises," "I ought not steal," and so on—it is just that for her the "ought" is a non-moral one. And, as I say, perhaps the moral error theorist will read Bloomfield's paper and believe it, thus arming herself with a foundation for thinking that self-interest will *always and for everyone* come out on the side of morality. None of this jars her commitment to a moral error theory in the least.

If any of this feels uncomfortable, then it may be useful to consider an analogy. Picture a theistic error theorist—better known as an "atheist." Suppose there were a kind of prescription that could be marked as "…according to God": "You ought not kill, according to God," "You ought not testify as a false witness against your neighbor, according to God," and so on. The atheist is unmoved by these prescriptions qua divine commands; he doesn't believe in God, so doesn't believe that there are any commands issuing from God, so does not believe that one ought not kill, according to God. It hardly follows, however, that the atheist is inclined to go around killing, or, indeed, that his reluctance to kill is any flimsier than that of the Pope. The atheist may be as determined to refrain from killing as anyone else, for any number of reasons. Perhaps he thinks that it is morally wrong, perhaps he has so much sympathy for his fellow humans that the thought sickens him, perhaps he recognizes certain forms of self-harm that would ensue from killing, perhaps all of the above. The atheist is still inclined to enthusiastically assert "I ought not kill"—and perhaps takes himself to have grounds for holding that this is true always and for everyone—but he will remain clear in his own mind that he is not employing the "ought…according to God" locution. And this, clearly, doesn't undermine his atheism in the least.

The comparison between atheism and moral error theory is useful to bear in mind when it comes to responding to a possible objection to what has been argued. The objection runs as follows:

> You error theorists argue that morality is flawed, yet you still think that we ought to refrain from stealing, keep promises, not initiate violence, and so on. But if the foundational moral question is "How ought one to live?" and you have answered this, by reference to self-interest, in such a way that the answer is "Keep promises, refrain from stealing, don't initiate violence, and so on" then you *have* endorsed a morality. You have allowed that moral normativity can be identified with prudence (or at least with a proper subset of prudence). So your moral error theory collapses.

[10] There is, of course, a kind of loyalty that is based on judgments of moral obligation. I submit, however, that one can also have *feelings* of loyalty—feelings of attachment and affection that involve desires to protect the welfare of another person—that need not be "moralized" by the subject in the least.

In order to understand the moral error theorist's response to this objection, some distinctions must first be drawn. We must note, to begin with, the sense in which even the moral error theorist "believes in morality": She believes that *moralities* exist, in the same way that the atheist recognizes that religions exist.[11] What the error theorist does not do is epistemically endorse any morality. I say "*epistemically* endorse" so as to exclude certain pragmatic ways in which a morality might be endorsed, such as approving of its practical output (agreeing that one ought not break promises, ought not steal, and so on), or acknowledging that the institutions of morality are instrumentally beneficial.

Note also that the error theorist need not have granted that there *is* a systematic answer to the question "How ought one to live?" Perhaps Ernie should live one way, given his circumstances and upbringing, and Bert should live another, given his. (Indeed, it may be precisely in virtue of thinking that there is no answer to the question "How ought *one* to live?" that someone is a moral error theorist.) But let us suppose that we are dealing with a kind of error theorist who, for whatever reason (perhaps having been convinced by Bloomfield's argument), accepts that the question can receive some kind of universal, systematic answer—that there is a way that "one" ought to live.

The above objection in fact suggests two challenges for this type of moral error theorist. The first is that acknowledging that the question "How ought one to live?" can receive any positive answer in itself constitutes or implies the epistemic endorsement of a morality. The second is that answering this question *in a way that underwrites a particular content* (keeping promises, not stealing, and so on) constitutes the epistemic endorsement of a morality. In both cases the moral error theorist will offer much the same answer: She will disagree because she believes that there is something special about moral normativity (something that, she thinks, is deeply flawed) such that merely to answer how one ought to live, or even to answer it in a way that underwrites keeping promises, and so forth, is insufficient to amount to the epistemic endorsement of a morality. Imagine, by analogy, the atheist facing the objection that insofar as he thinks that there is a way we ought to live then he is, despite himself, really a theist, *because that's all there is to theism*. He will, quite rightly, object that that's *not* all there is to theism, that to epistemically endorse a theistic framework requires subscribing to some substantive metaphysical theses about the existence of a divine being who enjoys such properties as omnipotence, omniscience, and so forth. It is in virtue of disbelieving these theism-constituting theses that the atheist is an atheist. Similarly, the moral error theorist also thinks that to endorse any moral system requires subscribing to some substantive (and, presumably, "metaphysical," in some broad sense of the word) theses, and it is in virtue of her disbelief in these theses that she is a moral error theorist. (The kind of theses in question were pointed to in the third paragraph of this essay.)

[11] I'll assume without argument that endorsing a religion entails endorsing theism. Some might object to this (raising the case of Buddhism, for example), but the niceties of that debate do not interest me.

In fact, the idea that giving *any* positive answer to the question "How ought one to live?" constitutes the epistemic endorsement of a morality seems highly implausible. Suppose the answer comes back: "Do whatever the hell you feel like." There would seem to be something terribly misleading in the insistence that living according to this rule constitutes endorsing a *morality*. (If one really wanted to stretch the word "morality" to this extent, the moral error theorist can always just disambiguate: "Well, okay, in that unnaturally strained sense of 'moral,' of course I endorse morality—but nevertheless there is a far more familiar customary usage regarding which I remain a disbeliever." We can imagine the atheist saying something comparable if faced with the serious assertion that *God is love*.) The objection, as it is stated previously, contains an element that implies that not just *any* positive answer to this question will constitute the endorsement of a morality; rather, there appears to be a contentful constraint on what can count: Prudence (or a proper subset thereof[12]) becomes a candidate for constituting a morality only to the extent that it endorses keeping promises, refraining from stealing, not initiating violence, and so on. But even with the addition of this constraint on content, the moral error theorist will—for the same reason as before—remain unimpressed with the proposal that she has, despite herself, endorsed a morality. Whatever argument or arguments have led her to embrace moral skepticism will almost certainly constitute grounds for resisting this objection. To repeat: The moral error theorist believes that for something to be morally bad (say) would require the instantiation of some property that (1) is not supplied by the universe (as a matter of fact or necessarily), and (2) is *essential* to moral badness, such that anything lacking this feature just won't count as moral badness. (For ease of reference, let us call this property the "special feature" that the error theorist attributes to morality.) Assuming that we are dealing with an error theorist who allows that there is nothing particularly fishy about prudence, then we are *ex hypothesi* dealing with someone who thinks that prudential normativity lacks the special feature that dooms moral normativity. Thus the moral error theorist will not think that prudence is a good contender for being identified with moral normativity: Someone whose deliberations are guided solely by prudential considerations—even if these considerations speak in favor of all the things that morality is typically assumed to prescribe—is not thereby epistemically endorsing a morality.

It might be objected—by a moral noncognitivist, for example—that morality is not the kind of thing that requires *epistemic* endorsement at all. It might be objected that the only kind of endorsement needed is practical, and that so long as a person is generally behaving himself then he is endorsing morality in the only sense that matters. However, the dialectical point that I am making is that whatever argument(s) have led a person to defend a moral error theory will include grounds for thinking that moral judgment is a matter of *belief*, that *epistemic* endorsement is coherent and called for. The objection under consideration is that the moral error theorist somehow undermines

[12] See note 2. For the sake of brevity I will drop this qualification about proper subsets.

her own position if she accepts prudential normativity and accepts that it speaks in favor of general niceness. This objection cannot be founded on an insistence that noncognitivism is true, for the error theorist *ex hypothesi* won't agree to this.

For all that, noncognitivism could be true; nothing I say in this essay is designed to convince anyone otherwise. At no point is my intention to establish that the moral error theorist is *correct*. Perhaps the special feature that the error theorist attributes to morality is instantiated by the universe after all. Or perhaps the error theorist is mistaken in thinking that this feature is an *essential* characteristic of moral normativity; perhaps a kind of normativity lacking this feature would nevertheless satisfy enough of our other desiderata to count as the real thing. A moral philosopher advocating an error theory must be prepared to defend herself on both fronts. This job is made difficult by the fact that it is often extremely difficult to articulate precisely what it is that is so troubling about morality. And this failure need not be due to a lack of clear thinking or imagination on the error theorist's part, for the thing that is troubling her may be that there is something deeply mysterious about morality. The moral error theorist may, for example, perceive that moral imperatives are imbued with a kind of mystical practical authority—a quality that, being mysterious, of course *cannot* be articulated in terms satisfactory to an analytic philosopher. Such an error theorist is forced to fall back on vague metaphors in presenting her case: Moral properties have a "to-be-pursuedness" to them (Mackie 1977: 40), moral facts would require that "the universe takes sides" (Burgess [1978] 2010: 13), moral believers are committed to "demands as real as trees and as authoritative as orders from headquarters" (Garner 1994: 61), the phenomenology of believing oneself morally required to act is to think "Well, I just *have* to" (Joyce 2001: 141), and so on. Indeed, it may be the very perniciously vague, equivocal, quasi-mystical, and/or ineliminably metaphorical imponderabilia of moral discourse that troubles the error theorist.[13] (For useful discussion of this point, see Hussain 2004.)

As I have indicated earlier in this essay, it is not my intention on this occasion to present any particular error-theoretic argument regarding morality. For a start, doing so would take too long, and, moreover, it is more useful here to keep things broad so as to give consideration to the moral error theorist in a generic sense (hence these unsatisfying references to a "special feature" that the error theorist attributes to morality). It might be thought that without presenting any particular argument it will be impossible to assess whether the error theorist is reasonable in claiming that prudential normativity cannot be identical to moral normativity. It might be thought that we really need to

[13] Compare Wittgenstein, who concluded that moral language is "nonsense" on the basis of his observation that moral discourse consists largely of similes, yet "a simile must be a simile for something … [but] as soon as we try to drop the simile and simply state the facts which stand behind it, we find there are no such facts" (Wittgenstein 1965: 10). Interestingly (in light of what I will discuss later in this essay), although he concludes that nonsense is "the very essence" of moral expressions (11), Wittgenstein adds that engaging in moral thinking is a tendency of the human mind that "I personally cannot help respecting deeply and I would not for my life ridicule it" (12).

have the error theorist spell out what she takes the essential and problematic feature of morality to be, so we can judge whether she is correct in claiming that prudence lacks it. But in fact I think that we can get a pretty good taste of how that argument will go without committing our (usefully generic) error theorist to any particular line of reasoning. Indeed, it seems to me that anybody—whether error theorist or not—should be extremely uncomfortable about any proposal to identify moral imperatives and values with prudential imperatives and values.

Let us begin by thinking about how prudential normativity works. Suppose it is claimed "Ernie ought not eat cookies in bed," using a plain and simple prudential "ought." The sentence is true (with the prudential "ought") only if eating cookies in bed will harm Ernie in some way. Perhaps doing so will lead to crumbs in his pajamas, leading to sleeplessness. But it is possible that there is harm to other parties involved too. Perhaps what is under consideration is Ernie's decision to eat cookies in *Bert's* bed, thereby annoying (harming) Bert, which will lead to Bert retaliating against (harming) Ernie in some way. (Or perhaps God punishes Ernie, or perhaps Ernie pollutes his own soul, or perhaps Ernie fails to respect himself, and so on—the details don't matter.) The important thing to notice about a prudential "ought" that involves harm to more than one party is the counterfactual asymmetry between the harms: If in eating cookies in Bert's bed Ernie will harm himself but somehow (magically, perhaps) Bert will escape harm, then the prudential claim would remain true; but if in eating cookies in Bert's bed Ernie will harm Bert but will somehow manage to avoid the self-harm, then the prudential claim would have to be retracted. (In the latter case, of course, it may remain true that Ernie ought not to eat cookies in Bert's bed, *using some other kind of "ought."*) Reflecting this, let us say that in prudential normativity the self-harm is *primary*—it is what *makes* the action imprudent.

Now let us contemplate the proposal that moral normativity might be identified with prudential normativity. (Note that I am not targeting the view that acting in morally bad ways is imprudent—Bloomfield's position—but rather the stronger identification claim that moral badness *is* imprudence.) Consider the Nazis, whose actions were so horrendous that even trotting them out endlessly as a philosophical example shouldn't dampen our horror at what they perpetrated. The error theorist may despise the Nazis as much as anyone, but nevertheless withholds assent from the claim that what they did was *morally* wrong. (Obviously, the error theorist needs to be careful in voicing this claim, for it is likely to be misconstrued as indicating some kind of tolerance for the Nazis, whereas in fact she simply thinks that all moral language is bankrupt: that the Nazis' actions were not morally wrong, not morally right, not morally permissible, not morally anything.) Let us focus our attention on a particular SS guard, who herded frightened Jewish children into the gas chambers with full knowledge of what he was doing. Let us stipulate that no possible defense could be mounted for his deeds; if any action is a moral crime, it is his.[14] Now let us adopt the proposal

[14] The moral error theorist who thinks that moral predicates have empty extensions across all possible worlds will struggle to take this last conditional phrase literally as a counterfactual truth. I submit, however,

that the wrongness of his actions is nothing more than their imprudence (that is, that moral badness is imprudence). This means accepting that what primarily makes the guard's action wrong is that he harmed *himself*. The fact that he harmed others contributes to the wrongness, but only derivatively (in that in harming others he harmed himself), and it is the harm to himself that really determines his wrongness. It also means accepting that what determines the *magnitude* of his crime is the magnitude of the injury he does himself (that is, in harming so many innocent victims he damaged himself severely). Furthermore, it means accepting a counterfactual: that if the guard had killed all those innocent people *but had managed somehow (magically, perhaps) to avoid the consequential self-harm*, then there would have been nothing wrong (that is, morally/prudentially wrong) with his actions.

This, I hope, sounds appalling. It might not be unreasonable for us to agree that the guard *did* harm himself in various ways, but the idea that the wrongness of his actions derives ultimately from that self-harm is a monstrous thought—almost as monstrous as the thoughts the guard uses to justify his actions to himself. The example illustrates the enormous difference between prudential and moral norms, and does so at an intuitive level, without pretending to articulate what a moral norm is. (We are supposed to think "Whatever exactly a moral norm is, it's not like *that*.") There are many ways to demonstrate the difference between these two types of normativity. To perform an action that harms oneself (for example, to drink strong coffee before going to bed) may amount to doing something that one ought not to do, but it's not the right kind of "ought-not-ness" to count as a *transgression*—and the notion of transgressing is surely fundamental to moral thinking. The "emotional profiles" of prudence and morality appear intuitively to be very different. Our basic emotional response to someone's self-harm is *pity*. The emotion of *retributive anger* makes little sense within the framework of prudential normativity, for what sense is there in the idea that someone who has harmed himself *deserves* the infliction of further harm (or, moreover, that the severity of the harm we inflict should be proportional to the degree of self-harm)?[15] Harming oneself per se doesn't (and shouldn't) provoke the emotion of *guilt*; it provokes the phenomenologically very different form of self-castigation of thinking "I'm so stupid" (and is *that* what we think the SS guard should be feeling?). Without underwriting guilt, it is implausible that prudential considerations could form the lifeblood of a moral *conscience* in the way that moral considerations do. Consider also the reparations that on many occasions we would insist that the moral criminal make to his victims. On the morality-qua-prudence view, the primary victim of any crime is always the criminal himself. Perhaps compensating the other victims (or simply apologizing to them) will be a means for the criminal to benefit himself, to undo the self-injury that

that even she can understand the spirit of the claim, and treat it as an acceptable rhetorical pronouncement that stands in for some true complex proposition.

[15] This is not to deny that there may be other grounds for punitive response for which the idea of *desert* plays no role. In punishing the SS guard we may hope to discourage him from harming himself in this manner again, or hope by example to discourage others from such heinous acts of self-harm.

he has inflicted, but there is no reason to assume that this is the only or the best way for him to accomplish this end, and thus if he finds some other way of compensating for the harm he did himself (taking a relaxing holiday?—treating himself to a special gift?—forgiving himself?) then this act of direct self-profit may well be the preferable course for him.

It may help to clarify my central claim—that moral badness and imprudence are non-identical—if it is observed that it is consistent with maintaining any or all of the following:

1. Performing actions of the types that are typically thought of as morally bad will cause self-harm.
2. (1) is true always, necessarily, and for everyone.
3. Moreover, the degree of self-harm is proportional to the magnitude of the (assumed) moral badness.
4. Some actions are considered both morally bad *and* imprudent, making it sometimes difficult to tease the two apart.
5. When we try to dissuade someone from performing a morally bad action, the negative consequences that will befall him are likely to be among the first things we mention. (We may even have a deeply entrenched and institutionalized cultural tradition of appealing to the punishments of an all-powerful divine entity in order to back up our moral judgments, thus ensuring that we think of moral transgressions as imprudent.)
6. Sometimes normative frameworks are "nested," such that one is obliged, according to framework A, to follow the prescriptions of framework B. Thus, in some circumstances a person may think it morally required to be prudent. This, again, makes it hard to tease the two apart, but does not indicate the absence of a distinction. (By analogy, a parental authority may decree to a child "Do what the teacher tells you to." If the teacher then orders "No talking," then we may say that not talking has been prescribed both directly by teacherly authority and indirectly by parental authority. But the two normative frameworks are nevertheless distinct, and their respective values and rules may have very different characteristics.)
7. Moral norms need not be exclusively other-regarding.[16] The sentence "You ought not neglect your health" may be used to express a piece of prudential advice, or could be used to state a self-regarding moral imperative. These respective usages would display different characteristics. (If used morally, for example, the "ought" claim will make legitimate certain kinds of criticism for non-compliance that a prudential usage would not.)
8. To observe the distinction between moral normativity and prudence is not to disparage prudence or suggest that it must take a backseat to morality.

[16] Compare Falk (this volume).

The form of argument pursued above—examining the characteristics of a normative system that is being offered as a candidate for vindicating morality, and declaring that it displays insufficient mesh with our pretheoretical desiderata concerning what moral normativity is like—is a regular task for the moral error theorist; she will find herself doing it again and again. Defeating the candidacy of *prudence* is fairly undemanding, I think, and can be successfully accomplished while keeping the discussion at a rough, intuitive level. But the error theorist's task may not always be so easy, and for other claimants it may be necessary for her to spell out in as much detail as possible what she takes to be distinctive (and problematic) about morality, analyze carefully the characteristics of the candidate, and compare the two. The error theorist may accept that some candidates fare better than others—some may have a much better claim than prudence—but she believes that ultimately none comes close enough to deserve the name "morality." The closest satisfiable satisfier of all our moral desiderata still counts at best as "schmorality."

Let me be clear what is meant by "schmorality" in this context. Picture a continuum comprised of what can be thought of (in a benignly vague manner) as "normative frameworks." At one end we have value systems that clearly count as *moralities*: Christian ethics, deontological systems, Moorean intuitionism, Platonic theories about the Form of the *Good*, and so on. The error theorist doesn't doubt that these moralities exist, but she thinks that none of them deserves to be epistemically endorsed. At the other end we have things that clearly don't count as moralities: the rules of chess, etiquette, doing whatever the hell you feel like, and so on. The moral error theorist is free to epistemically endorse the claims of such systems (for example, she thinks that "You must not move your knight in a straight line" is true). Somewhere on this continuum will lie normative frameworks for which it is not immediately apparent whether they count as moralities: Some people will think they do; others will think they don't. Call these items "contenders," of which one example is prudence. The error theorist, as we have seen, thinks that prudence is a poor contender for being a moral system. (Indeed, even those with no sympathies with moral skepticism should assent to this.) Note that calling prudence a poor contender for being a moral system is not to call it a poor moral system (which would imply that it is a moral system), any more than a hopeless contender for being elected president is thereby a hopeless president.[17] It is not that the error theorist fails to epistemically endorse prudence (she may agree that Ernie ought not eat cookies in bed); but rather she thinks—for the kinds of reasons outlined above—that there is simply insufficient mesh between prudential normativity and moral normativity for prudence to count literally as *a morality*. And the moral error theorist thinks this about *every* contender: Either it may be epistemically endorsed but is too far from the "morality" end of the spectrum to count literally as a morality, or it is close enough to count as a morality but (for various reasons) cannot be epistemically endorsed. Indeed, holding this combination of views is constitutive of being a moral

[17] Someone once claimed to me, in all seriousness, that golf was his religion. The correct response is not that golf is a very poor religion, but that it doesn't count literally as a religion at all.

error theorist. Every contender is thought to be either unsuccessful—that is, there is nothing in the world answering to its claims, there is nothing that renders these claims true—or a schmorality: something bearing a resemblance to a morality—enough, perhaps, to be mistaken for the real thing by the inattentive—but which falls short of really being so.

What determines whether something is a morality or a schmorality? In my opinion, the answer turns on how the concept *morality* is *used*. If concept A is used in a certain manner, but turns out to be problematic for various reasons (that is, it is uninstantiated by the world), and concept B is an instantiated contender for replacing A, then B can be an adequate successor only if it too can be used in the same manner. For example, even when we realized that nothing is absolutely simultaneous with anything else, the relativistic notion of simultaneity was able to take over seamlessly, since it works just as well in everyday contexts for creatures whose movements don't approach a significant fraction of the speed of light. We can *use* the concept of relative simultaneity in the same way as we can use absolute simultaneity, which suggests that the change didn't amount to replacing one concept with a different concept at all, but rather we just made a revision internal to a single concept. Thus we are not forced to the radical position that every pre-Einsteinian assertion of two events occurring simultaneously is false. By comparison, when we discovered that there are no diabolical supernatural forces in the universe, we had no further use for the concept *witch*. Perhaps we could have carried on applying the word "witch" to women who play a certain kind of local cultural role on the margins of formal society—perhaps we might even have located a cluster of naturalistic properties that all and only these women have—but carrying on in this way would not have allowed us to *use* the word "witch" for the purposes to which we had previously put it: to condemn these women for their evil magical influence and justify their being killed. Thus, there was little point in persisting in using the word "witch" to stand for certain instantiated naturalistic properties; we dropped it and concluded that all historical assertions that certain women were witches—even the loosely spoken ones—were false; we became error theorists about witches.[18]

The question, then, in the moral case, is "What do we use morality for?" The answer will almost certainly be extremely complex, and is, moreover, largely an empirical business. It is extraordinary how rarely this matter has been squarely faced, and deplorable that on those occasions that are exceptions, vague intuitions from the armchair have, more often than not, been thought to suffice. And yet on this question, as we have seen, depends the issue of whether all our moral utterances are true or false. If a contender for satisfying our pretheoretical desiderata for *morality* turns out to be something that we couldn't even use for the purposes that we have customarily put moral discourse—if, for example, we couldn't use it to justify deserved punishment, if it couldn't undergird the emotion of guilt, if it couldn't act as a bulwark against a range of motivational

[18] This paragraph is taken from Joyce (2006a: chapter 6).

infirmities—then we have good reason for thinking that we have in our hands but a *schmorality*. And if this is so of the *best* satisfiable candidate(s), then we should all be moral error theorists. Obviously, no deliberation of this kind can proceed until we know just what it is that we *do* use moral discourse and moral thinking for. Thus, until the jury delivers its verdict on this empirical matter, the fundamental metaethical disagreement between the moral error theorist and the moral success theorist (the cognitivist who believes that moral assertions are often true) remains at a stalemate.

Let me give one brief example of this kind of exchange, more for the sake of clarity than argumentative success. David Lewis offers a candidate for satisfying the noun "value": that "something of the appropriate category is a value if and only if we would be disposed, under ideal conditions, to value it" (1989: 113). The interesting details need not detain us here; the important point is that one of the discomforting implications of Lewis's offering is that, since human psychology is contingent, we might have valued different things (even under ideal conditions), thus there could have existed values different from those that actually do exist. Lewis's gentle example is that we might have valued seasickness and petty sleaze, but obviously far nastier things could have turned out to be good, according to his theory. Lewis admits that this rampant relativism is a disturbing implication, yet still thinks that his offering may be "as near right as we can get" to satisfying our problematic moral notions, supporting the conclusion that although "strictly speaking" the moral error theorist wins the day, "loosely speaking" values exist (137). Lewis may be correct. But how can we tell? How do we know when "Close enough is good enough"? According to my thinking, we must ask whether Lewis's "values" can play the same practical roles in our lives as moral values hitherto have done. What is interesting about Lewis's discussion is that he himself suggests a use to which we put values—one that turns out to undermine the candidacy of his favored claimant. The telling moment comes when he suggests why it is that relativism "feels wrong": He says that perhaps it is because "a large and memorable part of our discussion of values consists of browbeating and being browbeaten[19] . . . The rhetoric would fall flat if we kept in mind, all the while, that it is contingent how we are disposed to value" (135). Lewis's intention is to diagnose the source of our uneasiness about relativism, but if we take seriously the thought that such rhetorical impact is an important part of the use to which we put moral considerations (both interpersonally and, perhaps, intrapersonally), then he has provided us with evidence *against* the adequacy of his theory of value, since he has identified an important practical purpose that would be lost if we adopted his replacement concept. (It is perhaps a depressing thought that this might be a central function of moral discourse, but, as I declared above, this is something for which hopeful or romantic guesses won't stand in for evidence.[20]) Thus there is at least one consideration—by Lewis's own lights—in favor of thinking that

[19] Lewis here footnotes Hinckfuss (1987).

[20] Of the uses to which we put morality, to ignore some, in this calculation, on the grounds that they are considered "immoral" would, obviously, be to beg the question against the moral error theorist.

his "values" are not the real McCoy, in favor of thinking that he has provided us with a schmorality rather than a morality.

Suppose the error theorist is correct in holding that the closest satisfiable claimants for our moral concepts are all schmoral concepts. The question arises as to what she then does with moral concepts. The natural assumption is that the error theorist will also be an eliminativist: that she will recommend the abolition of moral language in all unembedded positive contexts. (These last qualifications are supposed to indicate that nobody thinks that we should eliminate moral language *altogether*; the error theorist will still assert things like "There exists nothing that is morally bad" and "St. Augustine believed that stealing pears was morally wrong.") The popular assumption is that if we catch a professed moral error theorist employing moral talk then we can triumphantly cry "Aha!" and accuse her of committing the intellectual vice that Quine (in a tone of disgust) characterized as engaging in "philosophical double talk which would repudiate an ontology while simultaneously enjoying its benefits" (1960: 242). Any such accusation is an argument not against the moral error theory but against the theorist—showing her to be a hypocrite, disingenuous, in bad faith, or vacillating between belief and disbelief. (Perhaps, on the latter charge, the error theorist is like Hume's Pyrrhonian, who, it will be recalled, cannot live his skepticism because "nature [is] too strong for it" ([1740] 1978: 657).)

But eliminativism does not follow logically from the error theory. The question of what one *ought to do* with one's moral discourse need not be a moral inquiry but may be construed as a practical question: Perhaps it involves a prudential "ought," or perhaps a hypothetical "ought" concerning how the agent's (idealized and fully informed?) desires may be optimally satisfied.[21] I do not intend to adjudicate on this matter; all that is of concern here is that it is a kind of practical question that (we have allowed) the moral error theorist has the resources to address. Let us just say that the error theorist will opt to eliminate moral discourse only if that conclusion is supported by some kind of cost-benefit analysis in comparison with other options. Yet what are the other possible options figuring in this calculation? The option of carrying on as if nothing has changed—of continuing to assert moral propositions and to hold moral beliefs *even while maintaining moral error-theoretic commitments*—is surely a non-starter, for the kind of doxastic schizophrenia involved in such a life not only violates epistemic norms but can also be expected to lead to various kinds of pragmatic handicap. But there is a third option: The error theorist may consider taking a fictionalist attitude toward morality. The fictionalist's point of departure is summed up nicely by Hans Vaihinger: "An ideal whose theoretical untruth or incorrectness, and therefore its falsity, is admitted, is not for that reason practically valueless and useless; for such an idea, in spite of its theoretical nullity, may have great practical importance" (1935, viii).

[21] These disjuncts are distinct on the assumption that psychological egoism is false (an assumption that I feel confident in making). The falsity of psychological egoism means that a person—even a moral skeptic—may have genuinely non-derivative desires for others' welfare. Any "ought" claim that constitutes advice on how such an altruistic desire will be best satisfied need not correspond to a prudential "ought."

To adopt a fictionalist stance toward morality is to continue to make moral *utterances* and have moral *thoughts*, but withhold assertoric force from the utterances and withhold doxastic assent from the thoughts. The fictionalist can be seen as an error theorist who attends to both epistemic and pragmatic norms.[22] His respect for epistemic norms means that he steadfastly refuses to *believe* any moral claim; his sensitivity to pragmatic norms means that he seeks and embraces the expedient option.[23] On the assumption that morality is in various respects useful when it is asserted and believed, eliminativism will (*ceteris paribus*) constitute a practical cost; and if morality is *very* useful then eliminativism will constitute a *big* cost. The fictionalist option, therefore, becomes attractive if (and only if) it promises to recoup some of these costs. The advocate of fictionalism holds that some of these losses may be recovered by adopting a policy of employing moral language, engaging in moral deliberation, and being moved by moral emotions, but throughout it all remaining disposed to deny the truth of any moral proposition if pressed in an appropriately serious manner (such as when in the philosophy classroom), thus not really believing any of it (thus not violating any epistemic norms), and thus deflating a host of well-thumbed philosophical problems concerning the ontology of moral facts and our access to them. Regarding actual moral discourse, the fictionalist remains an error theorist: He thinks that this discourse does aim at the truth but systematically fails to secure it. On the grounds of expediency he advocates a revolution in our attitudes toward morality, and regarding the (imaginary) post-revolution moral discourse, the fictionalist is no error theorist, for, come the revolution, moral discourse will no longer aim at the truth.[24] The tricky part of expounding fictionalism is to make out a kind of attitudinal acceptance *other than belief* that can play a central role in serious intellectual inquiry and serious practical deliberation.

There are many objections to fictionalism in general, and some to moral fictionalism in particular. (For discussion, see Hussain 2004; Kalderon 2005a, 2005b; Nolan, Restall, and West 2005.) In what remains I will discuss three objections that are similar in that each holds that moral fictionalism somehow undermines the error theory on whose shoulders it stands, thus rendering itself redundant (in the sense that if the

[22] Sometimes the label "fictionalist" refers to a philosopher advocating that we adopt a fictive stance; sometimes it refers to someone who has adopted that stance. (If certain critics of fictionalism are correct, there are no fictionalists in the latter sense.) Although potentially confusing, this equivocation seems benign in most contexts.

[23] I should like to draw attention again to the distinction observed in note 21. An error theorist may have reason to adopt the fictive attitude because doing so promises to satisfy certain of her (idealized and fully informed?) desires—and I see no grounds for denying that (some of) these desires may be genuinely altruistic in content. Thus, in fact, it need not be self-interest that recommends the adoption of the fictive stance. Nevertheless, counsel that appeals to self-interest is more likely to have a broader general influence, and thus (giving consideration also to the demands of concision) I will continue to fudge over this subtlety, and speak as if self-interest were the only relevant consideration motivating the fictionalist.

[24] The kind of fictionalism being described here is the "revolutionary" branch. In contrast, a "hermeneutic" fictionalist argues that we have been taking a fictive attitude toward the target discourse all along (and thus the hermeneutic fictionalist is not an error theorist). Hermeneutic moral fictionalism is advocated by Mark Kalderon (2005b) and criticized by Jason Stanley (2001).

error-theoretic account of Xs becomes implausible, then although taking a fictive attitude toward Xs remains an intelligible option, there is no need to do so). The first two objections can be interpreted as maintaining that anyone attempting to fictively accept morality must be epistemically endorsing a morality after all. The third objection doesn't quite amount to this, but is related in that it holds that embracing (a particular kind of) fictionalism will destabilize a particular kind of argument in favor of the moral error theory.

First, one might complain that if the policy of uttering and thinking moral propositions can be recommended on prudential grounds, then moral discourse has been vindicated after all. Indeed (the complaint might continue), the fictionalist has supplied evidence against his own error theory, since he has provided grounds for equating moral norms with prudential norms. This is somewhat different from the objection to moral error theories that we encountered earlier. Then the claim was that if the error theorist agrees that *acting* in accordance with assumed moral norms is justified on prudential grounds, then he has provided morality with all the justification that it needs. Now the claim is that if the error theorist agrees that *talking and thinking* in moral terms is justified on prudential grounds, then he has provided morality with all the justification that it needs.[25] But the response is much the same. We should start by bearing in mind the distinction between epistemic justification and instrumental justification. If someone holds a gun to your head and says "Utter the sentence '1 + 1 = 3' or I'll shoot!" then the act of utterance will be prudentially wise (instrumentally justified), but the content of the utterance will be no less false—and any act of believing it no less illegitimate (no more epistemically justified)—for that. Recall that the error theorist has been impressed by the thought that moral propositions have substantive metaphysical (and problematic) implications or presuppositions that prudential propositions lack. The fact that the act of uttering one of these flawed sentences may be instrumentally justified hardly shows that the sentence must be true, or that believing the sentence is epistemically justified. Nor does the fact that uttering a normative sentence is prudentially justified mean that the sentence really expresses nothing other than a prudential norm. On this last point it might be useful to consider a comparison. A person might choose to cultivate the personality trait (assuming that it deserves to be so-called) of having altruistic emotions toward his friends and family.[26] Quite how one goes about such an act of "cultivation" need not bother us now; the point is that some act of deliberate choice is involved, which, if successful, results (at some time in the future) in having interests in the welfare of certain others—interests that do not depend on the contribution that the others' welfare makes to one's own interests. The important thing

[25] Crispin Wright may be interpreted as presenting an argument along these lines. See Wright (1992, 1996).

[26] Note that here I am considering altruistic emotions in a non-moral sense. To *like* someone—to have a non-derivative concern for his welfare, to be motivated to act to further his interests, to feel affection toward him—is a capacity that might be enjoyed by a creature entirely lacking the cognitive sophistication to make any moral judgments at all. (See notes 9 and 10.)

to notice is that at the time of original deliberation the person may be calculating entirely in selfish terms; she may realize that having altruistic emotions will, in various ways, contribute to her own welfare. This observation, however, in no way undermines the possibility that the love and sympathy that this person eventually comes to feel are genuinely altruistic in nature. One can be selfishly motivated to become a less selfish person, and may succeed. Similarly, one can be motivated on grounds of self-interest to adopt a policy of accepting a certain class of normative claims—which are distinct from prudential claims—and may succeed.[27]

The objection just discussed was that moral fictionalism undercuts its own error-theoretic basis—that adopting a fictive attitude toward morality amounts to an epistemic endorsement of it—and thus if one wants to maintain a moral error theory one had better eschew fictionalism, which more or less amounts to advocating that the error theorist be an eliminativist. The second objection is that the fictionalist stance is incoherent because the distinction between belief and "acceptance" cannot be maintained (see Putnam 1971: 68–9; Newman 1981). On this view, if someone acts, talks, thinks, and feels in accordance with having moral beliefs, then he actually *does* have moral beliefs. Thus, this objection also amounts to the allegation that attempting to adopt a fictive attitude (about anything this time, not just morality) will amount to an epistemic endorsement, and that if one wants to be an error theorist one had thus better steer clear of fictionalism.

Since *belief* is a contested notion, the suspicion arises that some accounts of belief will allow for a separate category of *acceptance* while others—for example, neo-behaviorism—will not. And so it may seem that the only means of responding to this objection is to provide a convincing argument for one of the former accounts. But in fact there is good reason for thinking that *all* parties have cause to allow this distinction, even the neo-behaviorist. Consider the crudest kind of behaviorism that says that all it is to believe that p is to act as if one takes p to be the case. Even so boorishly extreme a behaviorism will want to allow that on occasions a person may act as if she takes p to be the case without believing that p. Actors, for example. This observation alone forces the acknowledgment of some category of acceptance distinct from belief: It is the attitude actors take toward elements of the fiction into which they enter. The thing about actors, of course, is that they are disposed to "step out" of the fiction; they don't act *all the time* as if they take p to be the case. But the crucial detail to notice about the fictionalist is that he too remains disposed to step out. There are contexts where he does *not* speak as if he takes p to be the case: namely, when he is in the critical context of declaring his endorsement of the moral error theory.

But the person pressing this objection may persist. Even though acknowledging some kind of attitude—distinct from belief—deserving the name "acceptance," she may doubt that one can be in this state with respect to some subject matter *nearly all the time*. Our crude behaviorist may revise slightly: All it is to believe that p is to act *at*

[27] Compare Schmidtz (1994), on "maieutic ends."

least nearly all the time as if one takes *p* to be the case. The idea that the matter might depend on the *amount of time* one spends "immersed" in the fiction compared to the amount of time one spends "outside" it, strikes me as terribly improbable. What constitutes "nearly" here? Even acknowledging that the answer may be vague ("Around 90 percent"), it seems crazy to think that if I spend 95 percent of my time acting as if *p* were the case then I believe that *p*, but if I spend only 85 percent of my time acting in this way—all else remaining the same—then I do not believe that *p*. (I'm sure that there's a better objection to this than "Horrible theory!" but I'm happy on this occasion to rest matters there and trust in the reader's agreement.)

Far preferable would be the provision of some account of the *nature* of the two kinds of context, such that we can see that in one context utterances match what one really believes, even if it is a context entered into very rarely in comparison with the other context. Previously, I called the context of expressing disbelief (for example, when doing metaethics) the more "critical" context, and this is the term I have used on other occasions (Joyce 2001, 2005). It is, perhaps, an ill-chosen word, since it suggests that there is something "uncritical" about the fictionalist's engagement with moral matters in everyday life. It is important to see that "critical" here is a term of art, indicating an asymmetrical relationship between the two kinds of context (or, rather, naming a pole at one end of a continuum of contexts). Context *n* is more critical than context *m* if and only if *n* is characterized by a tendency to scrutinize and challenge the presuppositions of *m*, but not vice versa. This is consistent with *m* being the more "critical" in a vernacular sense of the word. For example, working out the plot of a complex novel may involve a great deal of careful thinking, whereas the thought "It's all just a fiction" is a simple matter. Nevertheless, in the sense intended, the latter is the more "critical" context since it questions the world of the novel. Similarly, when immersed in morality the fictionalizing error theorist may deliberate extremely carefully about consequences, weigh outcomes thoroughly, deploy acute powers of imagination and reflection, and so on, and yet still not inhabit his "most critical context" where he denies moral truth across the board. Although this amounts to not much more than a gesture—most prominently leaving us wondering just what is meant by "scrutinize and challenge"—I believe it is a promising way of addressing the problem, which, if successful, will make the *amount of time* one spends in the critical context irrelevant to the question of what one believes.

Those who doubt the viability of the belief/acceptance distinction may have their skepticism alleviated if they reflect on the seeming ubiquity of the phenomenon—or, at least, of closely related phenomena. The human proclivity for engaging with fiction (novels, movies, and so on) is the most conspicuous example, but arguably there are many less obvious instances of similar mechanisms operating in everyday life. Michael Bratman (1992) has argued persuasively that all practical reasoning involves accepting (but not believing) certain propositions as a background to effective deliberation. On the assumption that psychological simulation involves a kind of acceptance-without-belief, acceptance may be implicated in hypothetico-deductive reasoning, ascribing

mental states to others, and predicting others' behavior (see discussion in Davies and Stone 1995a, 1995b). Simulation probably plays a central role in empathy (Goldman 1992) and visual imagination (Currie 1995). Vaihinger (1935) supplies numerous mundane examples of our treating something "as if" it were true while knowing that it is not. (While Vaihinger almost certainly errs on the side of over-enthusiasm, his catalog of examples of the fictional stance is nonetheless instructive.[28])

The fictionalist's strategy here is unashamedly one of finding partners in innocence. Although it is unlikely that there is a single belief/acceptance distinction that all the aforementioned phenomena exhibit, there is enough family resemblance here that it is not unreasonable for the fictionalist to think that by cozying up *his* kind of belief/acceptance distinction to these other commonplace examples he can dispel some knee-jerk doubts. Is someone who reads a novel disingenuous or self-deceived? Is someone who engages in role-play suffering from anything deserving the name "schizophrenia"? Is someone who accepts a proposition as a background assumption when deliberating manifesting bad faith? Does feeling empathy make one a hypocrite? And does engaging in any or all of these practices have deleterious effects on one's interests? I take it that the answer to all is "No." Now, as admitted, the kind of belief/acceptance distinction at the heart of the moral fictionalist's case may not be quite the same as these other instances, but its similarity to these "innocent" examples is sufficient at least to show that such accusations (that the practicing fictionalist is in bad faith, suffers from self-deception, and so on) cannot be pronounced lightly. The onus, of course, is first on the fictionalist to articulate with precision what the distinction he has in mind amounts to; accusations of bad faith, schizophrenia, and so forth must be suspended until then. Then, of course, the burden falls to the opponent of fictionalism to replace the vague rhetorical sense of terms like "bad faith" and "schizophrenia" with something literal (and obviously undesirable).

The third objection to fictionalism that I will briefly comment on targets a particular brand of moral fictionalism—but since it is a kind that I have on occasion defended (Joyce 2001, 2005) I feel moved to respond. The fictionalism in question is one that hypothesizes that engaging in moral discourse is useful in a particular way: namely, that this engagement stimulates *motivation* in a pragmatically desirable manner. (Any fictionalist theory that assigns a different sort of usefulness to morality will not be affected by this objection.) It may be hypothesized, for example, that the expediency of moral discourse derives from its capacity to act as a bulwark against various kinds of practical infirmity—for example, weakness of will, discounting future gains, and so on—better than clearheaded instrumental deliberation. Thinking of an action as something that "just *must* be done" may encourage performance of that action more

[28] There are other "belief versus acceptance" distinctions in the philosophical literature that probably have little to do with the phenomenon (or family of phenomena) that is relevant here. Bayesian decision theorists often distinguish between partial belief and full acceptance (see Swain 1970). There is debate about whether collectives of individuals can have belief or merely acceptance (see Wray 2001). See also Cohen (1992); Frankish (1998).

reliably than explicitly conceiving of the action as one that serves one's long-term best interests; imagining the omission of that action to be something that will not merely frustrate one's desires but make one *reprehensible* and *deserving of punishment* may be more likely to result in resolve to perform the action. This, it seems to me, is an intuitively attractive idea, especially when it is made clear that the moral judgment may come "embedded" in an emotion, such as guilt or punitive anger. There is plenty of empirical evidence that self-directed moral emotions have motivational efficacy (see Carlsmith and Gross 1969; Freedman 1970; Tangney and Dearing 2002; Ketelaar and Au 2003; Zhong and Liljenquist 2006; Tangney et al. 2007).

So what problem does this hypothesis pose for fictionalism? There is of course the burning question of how taking a *fictive* attitude toward a set of norms and values could possibly engage motivation in this way. But that is an empirical question that I do not propose to discuss here (see Joyce 2001, 2005, 2006a); rather, I am interested in the theoretical question of whether supporting this hypothesis makes trouble for the error-theoretic basis of fictionalism. There are two reasons for thinking that it might. The first is that if a moral judgment engages motivation in this manner, then doubt is cast on the claim that moral judgments are a cognitive affair. But if moral judgments are in fact a *noncognitive* affair, then the moral error theory collapses, for one of the distinguishing features of this metaethical theory is its commitment to cognitivism.

This objection is confused. Noncognitivism is a thesis about what kind of mental state(s) moral judgments express; it denies that the state expressed is *belief* (that is, it denies that moral judgments are assertions). One popular form of noncognitivism—emotivism—claims that what is expressed is some (specifiable) conative or emotional state. To advocate cognitivism, however, is not to make the wild claim that moral judgments have nothing to do with emotions. Cognitivism is compatible with the view that moral judgments reliably prompt emotional activity. It is compatible with the view that moral judgments generally, or even always, flow from seething emotional activity in the brain. It is compatible with the view that what goes on when one makes a moral judgment is that one "projects" one's emotional life onto the events of the world. It is compatible with the view that the human capacity for moral judgment is a discrete biological adaptation that evolved precisely by virtue of its tendency to affect human emotions in a fitness-enhancing manner. None of these possibilities—nor, indeed, all of them jointly—entails the denial of the claim that moral judgments are assertions. Cognitivism is compatible even with the claim that the connection between moral judgments and emotional activity is a *necessary* one (though I should add that this is not a claim I endorse). Consider, by analogy, the act of promising. The criteria for a promise to have occurred involve a range of linguistic conventions in which both promiser and promisee need be versed; for example, the addressee must hear and understand the words uttered, and the speaker must take it that this is the case. If the addressee doesn't hear the "I promise . . " claim, or doesn't understand what the utterance signifies, then the act of promising misfires, and no promise has occurred (see Austin 1962). The satisfaction of these criteria will require both speaker and addressee

to have certain *beliefs*—for example, the speaker must believe that his addressee hears and understands. This connection is a necessary one: It is not possible that any person could succeed in making a promise to another person without having such a belief. And yet we would hardly say that the act of promising *functions to express the belief that one's audience hears and understands* (rather, a promise expresses an intention). Therefore, since a kind of speech act and a mental state may be necessarily linked without the former *functioning to express* the latter, then even if it were the case that moral judgments necessarily engage motivational states, noncognitivism is not the automatic conclusion (see Joyce 2002).

The second potentially problematic implication of assuming that the usefulness of moral thinking lies in its impact on motivation is even more limited in scope: It is a problem only for the error theorist who has employed a particular kind of argument to establish her error theory. Several philosophers who harbor skeptical misgivings about morality derive their doubts (in part) from a commitment to a Humean psychology, according to which beliefs and desires are distinct and only contingently linked states (see Williams 1981; Mackie 1977; Joyce 2001). John Mackie, for example, thinks that moral imperatives imply external reasons claims[29] (to import Bernard Williams's terminology), but, like Williams, he thinks that all non-institutional reasons claims are internal. (An internal reason is one that suitably connects with a person's "subjective motivational set"; an external reason is one that does not.) The basis of this latter opinion (for Williams at least) is the thought that any reason must potentially motivate a person, but only internal reasons can do so—and his ground for thinking this is that believing oneself to have an external reason cannot (alone) prompt motivation, because no belief can do that.

It should be stressed again that one might be a moral error theorist on grounds having nothing to do with any of these considerations. But if one *is* moved by this argumentative thread to adopt a moral error theory, and if, in addition, one is moved to become a fictionalist by the thought that morality is useful *because of its influence on motivation*, then, it may be supposed, one has some explaining to do. (It may sound like a fairly specialized position that is being defended—and indeed it is so—but it is not an uninhabited position, and the objection has been raised on more than one occasion.[30]) How could a moral *belief* (understood in Humean terms) function to influence motivation? In fact, the explaining is fairly easy. The hypothesis that moral judgments are useful because they influence motivation need involve only the claim that moral judgments *often* or *reliably* or *defeasibly* or *contingently* engage motivational structures. (Indeed, even "sometimes" will suffice.) It is perfectly possible that moral beliefs are just that—beliefs—and that beliefs alone never prompt motivation; but it may also be claimed that such beliefs, when they figure in an ordinary person's psychological economy—an

[29] See Mackie (1982: 115).

[30] Yes, I'm afraid I've been reading the reviews of my own book (Joyce 2001); see McKeever (2003); Wallace (2003). See also Hussain (2004).

economy that includes typical desires and emotions—will generally have an impact on motivation.

Having deflected these criticisms (and finding no other compelling), I feel confident in claiming that the fictionalist position is at least *coherent*. Whether the fictionalist stance is psychologically feasible, and whether it will supply the promised pragmatic gains, remain serious empirical uncertainties. Although on other occasions I have advocated the case for fictionalism (Joyce 2001, 2005), it must be underscored—as I did at the time—that while there is certainly a place for plausible speculation when it comes to directing people's attention, nothing confident can be claimed in advance of the *a posteriori* footwork. Perhaps in the end the data will not favor the fictionalist option. Perhaps eliminativism will be the better course for the moral error theorist, in which case she may use the term "schmorality" in its customary pejorative sense: to scoff "*Morality, schmorality*!" But it is important to remind ourselves that even the eliminativist error theorist will still have plenty of good and strong reasons—many of them self-interested reasons—for being nice to her fellows.

3

The Accidental Error Theorist

Introduction

The moral error theorist holds that morality is flawed in the same way as the atheist holds that religion is flawed: The discourse aims at the truth but systematically fails to secure it. Of the various possible arguments that might lead one to this form of moral skepticism, John Mackie's are the best known and most influential (Mackie 1946, 1977, 1980). Mackie's skeptical case targets both moral prescriptions and moral values, and in both cases it is their putative association with a kind of *objectivity* that bothers him. He is not an error theorist about prescriptions and values per se; he is always careful to identify the target of his doubt as "*objective* prescriptions" and "*objective* values." It is possible that one might be a moral error theorist for entirely different reasons (see Joyce 2007b: 52, 2011b), but here we will stick with Mackie's grounds.

The moral error theorist's opponents can be divided broadly into three types—one of which will be the focus of this essay. First, some opponents are noncognitivists, who deny that moral discourse aims at the truth. For all those remaining opponents we can use the label "success theorists"—those who both embrace cognitivism and hold that moral discourse often succeeds in hitting the truth.[1] There are many ways of taxonomizing success theorists, but for present purposes—clarifying types of opposition to Mackie's error theory—they can be divided into two. Some success theorists respond to Mackie's argument *head-on*: They argue that the troublesome concepts *objective prescription* and *objective value* are, when properly understood, perfectly defensible. Other success theorists pursue a *concessive* strategy: They concede that Mackie's target concepts are indeed philosophically indefensible, but they nevertheless resist the error theory by maintaining that substantial tracts of moral discourse remain unscathed by the skeptical critique.

Were this essay concerned with the head-on response, we would proceed to discuss what Mackie means by "objective prescription" and "objective value," investigate the merits of his famous Argument from Queerness against such phenomena, and so forth. But given that this essay is concerned solely with the concessive strategy, many of the details of that discussion can be put aside. The concessive success theorist and the error theorist *agree* that objective prescriptions and objective values are too queer to be admitted into our ontology, and since the present essay takes this agreement as its

[1] The term "success theory" comes from Sayre-McCord (1986).

point of departure, we need not pause to wonder about its precise content or its reasonableness. (This focus should not be taken to imply that I take a dismissive attitude toward noncognitivism or the head-on strategy against the error theory—though I do confess to thinking that the concessive route is the most promising.)

The concessive opponent will allow that *if* morality consisted only of objective prescriptions and/or objective values, then morality should indeed go the way of phlogiston and astrology. But, he insists, the antecedent doesn't hold. First, it may be argued that a notion like *objectivity* is just too recherché to be attributed to vernacular moral concepts; everyday moral discourse is committed to nothing of the sort. Second, it may be maintained that even if references to objective prescriptions and objective values are distinctly present in ordinary moral thinking, these are just two elements among others; the problematic elements could be discarded, while leaving us with a perfectly serviceable and unflawed moral system. Third, it may be pointed out that even if our actual morality consisted entirely of these elements, it would not follow that they are an *essential* feature of morality; again, after excising the defective elements, we would be left with a robust normative system still deserving of the name "morality." Hence, from the concessive strategist, the error theorist faces three kinds of accusation of narrow-mindedness: She is criticized for projecting her own philosophical extravagances onto the ordinary folk; she is charged with seeing only a part of morality and assuming it to be the whole of morality; and she is accused of seeing only the contemporary Western moral tradition and assuming that morality *must* be this way.

These accusations of narrow-mindedness cannot stand unsupported, however. It is incumbent on the concessive opponent of the error theory to identify those elements of normativity that (a) survive the skeptical critique and (b) are sufficient to constitute a morality. In other words, the viability of the concessive strategy depends on the viability of the positive theory on offer. It is accepted by both relevant parties that the offering must not imply the existence of objective prescriptions or objective values (at least of the kind targeted by Mackie); the principal question is whether it can be recognizably a *moral* system.

In this essay I will identify and critically assess several metaethical theories that can be interpreted as offering a concessive response to Mackie's error theory. The first perspective is the dispositional view of moral properties, the discussion of several versions of which will comprise the bulk of this essay. The dispositional theorist can be interpreted as agreeing with Mackie that objective prescriptions and objective values are too weird to be philosophically tolerated, and as responding that we should therefore identify moral properties with a certain class of *non-objective* prescriptions and values. The second perspective is virtue ethics, which will be discussed more briefly toward the end of this essay. The virtue ethicist can be interpreted (though with some strain, I admit) as agreeing with Mackie that objective prescriptions and objective values are too weird to be philosophically tolerated, and as responding that we should therefore begin our ethical inquiry not trying to make sense of prescriptions and values, but rather begin with normative questions about *character*.

This essay is not so ambitious as to try to show that these programs cannot succeed—I will not get even close to that conclusion—but I do want to show that neither side-steps the error theory as easily as is sometimes thought. In particular, I shall argue that proponents of these views are often playing directly into the hands of the moral skeptic by unwittingly championing error-theoretic views.

How is it possible to be an *accidental* error theorist? If Mackie is correct, then everyone who participates sincerely in moral discourse (which is, presumably, the vast majority of speakers) makes commitments that render the moral error theory true; and since speakers do not intend for this to be the case, there is some sense of "accidentality" involved. Yet to have the error theory be true of you—or (speaking more carefully) true of a discourse in which you engage—is not to *be* an error theorist. To be an error theorist is to take a metaethical stance; it is to offer a concrete opinion on the nature of moral discourse. But how could one perform such a purposeful and reflective action *accidentally*? What I have in mind is that some moral philosophers offer metaethical views that are intended to be versions of success theory but which, unwittingly, commit them to an error theory. To give a simple example, suppose that a metaethicist defends a divine command theory (DCT): identifying the moral property of *moral obligation*, say, with whatever is willed by God. Such a person presumably will be a theist. But suppose that in fact atheism is correct and there are no gods. Thus, since the predicate "…is willed by God" has an empty extension, then so too will the predicate "…is morally obligated." And if we assume that this is so of all other moral predicates too, then all sentences of the form "X is M" (where M is any moral predicate and X names something actual) will be false. The divine command theorist will have purposefully and reflectively asserted a metaethical view, but one which inadvertently commits him to an error theory.

Notice that it is not just any old problem with the DCT that would lead to this result. Most traditional objections to the theory can be categorized as casting doubt on the acceptability of identifying moral properties with properties pertaining to God's will. If one or more such objections were to prove entirely victorious, then although the divine command theorist would be shown to be in error, he would not thereby be shown to be *an error theorist*. By contrast, the result upon which I am focusing depends crucially on that identity claim (or weaker biconditional, as the case may be) being accepted and embraced (if only for the sake of argument), and then the failure of the right-hand side implies an inadvertent error-theoretic commitment.

Notice also that I am ignoring a potentially interesting distinction between those divine command theorists who, were they to come to believe that the right-hand side fails, would nevertheless maintain the DCT identity claim (thus acquiescing to the error theory), and those who would rather retreat from or revise this claim (thus avoiding the error theory). Those with the former disposition are in some sense more strongly "committed" to the error theory than those with the latter disposition. However, I propose to disregard this distinction; my interest here is not in how the proponent of the DCT would respond upon coming to believe in the failure of the

right-hand side. Perhaps the existence of God's will is so strongly an article of faith for a given person that nothing could bring her to believe that the right-hand side suffers from failure of reference. Such a person may be committed to the error theory not in the sense that anything could get her to acknowledge the error theory, but in the sense that she advocates a claim from which it follows (it is at least reasonable to suspect, even if *she* cannot be brought to suspect it) that no moral predicate is instantiated.

The dialectic can be simplified as follows. Suppose Ernie accepts proposition P but rejects proposition Q. Bert, by contrast, supports Q and also believes proposition R, which, when combined with P, implies Q. Bert accuses Ernie of being committed to Q. Ernie's natural reaction is to deny R, and perhaps nothing will budge him from this denial. If Ernie could be persuaded to accept R, he would face a choice: He could grudgingly admit Q or he could retreat from P. (We'll assume that denying "If P&R then Q" is unacceptable.) It might be protested that if Ernie's disposition is the latter—to reject Q by rejecting P—then he is not *really* committed to Q at all. My attention, however, is not directed at Ernie's possible reactions, but rather at Bert when he advocates R and observes its implications. Bert could, of course, simply argue for R; perhaps this is the real nub of their disagreement. But supposing that Bert is aware that Ernie's whole motivation for embracing P was to avoid Q—indeed, Bert has suffered Ernie's objections to Q which take the form "Not Q, because P instead!"—then we could hardly fault Bert for taking glee in defending himself by pointing out that by embracing P, Ernie has "inadvertently" committed himself to Q. (That there may be an even stronger kind of commitment Ernie might have to Q—that he is disposed, upon accepting "If P&R then Q," to grudgingly consent to Q—is another matter.)

Similarly, the divine command theorist will presumably believe that the predicate "…is willed by God" has a non-empty extension, and if we believe otherwise then this could be the topic of debate. But a more playfully provocative way of voicing our objection would be to accuse the divine command theorist—a would-be success theorist—of being an inadvertent advocate of moral skepticism. (That this is provocative is revealed by one's temptation to add an exclamation mark; that it is playful is revealed by the fact that one is disposed to deliver the objection with a grin.) My contention is that this happens in metaethics more often than is generally acknowledged. The number of unwitting moral error theorists is probably larger than the number of witting moral error theorists.

Response Dependent Morality

The advocate of a response dependency (RD) account of morality is a conspicuous example of the kind of opponent of the error theorist that has been under discussion. The RD theorist likely agrees with Mackie that objective normativity would be unacceptably odd—if, at least, by "objective" we mean something like (when applied to properties): *possessed irrespective of anybody's attitudes or psychological responses*. But the RD theorist is unimpressed with Mackie's attempts to convince us that any such

robust objectivism "has been incorporated in the basic, conventional, meanings of moral terms" (Mackie 1977: 35). The RD theorist argues that there is nothing unacceptable in the idea of a *non-objective* morality—a morality that is, in some manner (which Mackie would deny), constituted by our psychological responses to the world; a morality that we *make*.

When it comes to the RD theorist's positive proposals, there is a great deal of variation, and my intention is not to criticize the program *tout court*, but rather to point out some pitfalls along well-traveled paths. My target will be restricted to versions that make central the idea of a *disposition*. There are numerous versions of metaethical dispositionalism, many of which do not fall foul of the problems I will raise. The ones that do fall foul I will divide into two: negligent dispositionalism and optimistic dispositionalism. These terms will be explained shortly.

According to the dispositionalist RD theorist, moral properties are to be identified with dispositional properties, where the dispositions in question concern the generation of some kind of psychological response. The standard equation is this:

Moral goodness = the disposition to produce R in S in C

where "R" denotes a psychological response, "S" a type of subject, and "C" a set of circumstances (and where each of the three variables can be specified independently of the others). (See Johnston 1989, 1993; Lewis 1989; Casati and Tappolet 1998.) The substantive variation among different versions of dispositionalism arises from the different ways in which these three variables might be filled in (and the logical relation between the two sides).

Although I am classifying such theories as conceding a retreat from moral *objectivity*, there nevertheless remains a sense in which such properties retain a kind of objectivity. The disposition to produce R in S in C might be instantiated in an object even if there are no minds in existence (no Ss having R), and thus goodness, on this model, would be existentially mind-independent. (See Pettit 1991.) However, the model renders goodness non-objective in at least some other sense: The concept of the disposition to produce R in S in C cannot be articulated without making reference to a mental event (R), and thus goodness remains conceptually mind-dependent. To the extent that "response independent" is legitimately used as a synonym for one kind of "objectivity" (a kind that might be associated with primary as opposed to secondary qualities), it can hardly be denied that RD theorists embrace some kind of non-objective morality. We can make the simplifying assumption that whatever kind of moral objectivity RD theorists reject is precisely the kind that Mackie thinks is an essential but problematic aspect of moral discourse. (Recall that Mackie likens moral phenomenology to the perception of primary qualities: 1980: 34.) These theorists concur with Mackie that *that* kind of moral objectivity is unavailable.[2]

[2] The possibility of different kinds of objectivity reveals a shortcoming in my earlier classification of cognitivist opponents of the moral error theory into *head-on* versus *concessive* strategists. Suppose that Mackie successfully refutes a certain kind of objectivity for morality—we'll call it "type-A objectivity."

How could a proponent of dispositionalism be an accidental error theorist? Simply this: If the descriptive phrase on the right-hand side of the equation fails to denote a property, or denotes a property that is uninstantiated in the actual world.[3] We must, of course, be careful to distinguish an uninstantiated dispositional property from a non-manifest dispositional property. Consider the disposition to squeal if kicked unexpectedly. This dispositional property might be instantiated by something—a small dog, say—even if the creature is never kicked and never squeals; even if, that is, the disposition never becomes manifest. I am not looking for non-manifest dispositional properties; I am looking for descriptive phrases that purport to denote actually instantiated dispositional properties (whether manifest or not) but fail to do so.

One way that this might occur is when the descriptive phrase is incomplete. Consider the phrase "the disposition to squeal." Nothing has this disposition—nothing *can* have this disposition—for no disposition is picked out; the description is only partial. One no more succeeds in picking out a property with the phrase "the disposition to squeal" than one would succeed in denoting an object using the partial definite description "the book that is between the."

For completeness, a dispositional description needs to specify a stimulus event (for example, being kicked unexpectedly), a manifestation event (such as squealing), and conditions of stimulus. Often the last item can be specified tacitly. We could point at a particular small dog and ask "Would this dog squeal if I kicked it?" No circumstances are mentioned, but they are nevertheless implied: The question might assume that we are referring to the circumstances that the dog is actually in as we point at him, or assume, albeit vaguely, that we are referring to the "typical" circumstances in which one might encounter this small dog (thus excluding circumstances where there is no oxygen present, where the dog is exhausted from already being kicked, where the dog is wearing a little suit of armor, and so on). The latter might trump the former. We might point to a particular dog that is wearing a little suit of armor, acknowledge that if we were to kick it here and now it would not squeal, but nevertheless maintain that it has the disposition to squeal if kicked, inasmuch as it would squeal if kicked in *ordinary* (sans armor) circumstances.[4]

Suppose that there is at least one other kind of objectivity possible: type-B. Someone might start out pursuing a concessive strategy: agreeing with Mackie that there are no such things as objective [type-A] prescriptions, while nevertheless denying that type-A objectivity is an essential feature of morality. But now suppose that this same person goes on to defend the existence of objective [type-B] prescriptions. We could interpret this now as an instance of a head-on strategy: maintaining that there *are* such things as objective prescriptions, while insisting that Mackie has misconstrued their nature.

[3] Sometimes Mackie is interpreted as claiming that there is something *incoherent* about moral predicates, such that the error theory holds *necessarily*. I do not think this is a correct reading of his position (see Joyce and Kirchin 2010: xvi), but in any case, irrespective of Mackie's views on the matter, the most natural characterization of the moral error theory will allow that holding moral properties merely to be *actually* uninstantiated suffices to satisfy the criteria.

[4] It should also be noted that there can be a certain arbitrariness as to whether aspects of the disposition are specified as elements of R or S or C. Consider, for example, the trait of full information. We might speak most naturally of the disposition to produce R in *fully informed Ss* in C. But we could instead pack the

Now let us turn to moral dispositionalists who neglect to specify some elements of the disposition to such an extent that their dispositional phrase in fact fails to denote any property. Suppose moral goodness were identified with the disposition to produce approval in observers. I will assume that "approval" is an attitude that can be adequately specified, so will assume that the variable R is unproblematic. When the type of subject (S) is simply "observers," then the immediate question prompted is "*Which* observers?" Literally anything could produce approval in *some* observers (after all, we haven't yet restricted ourselves even to *human* observers), and I take it that any analysis of moral goodness that renders literally everything morally good can be rejected. One solution is to restrict the type of observer that is relevant; another solution is to modify the account in a relativistic direction (where "*o*" ranges over observers):

($\forall_o$) Moral goodness (for *o*) = the disposition to produce approval in *o* in C.

The latter solution might have the consequence that everything is morally good *relative to someone*, which is not quite so unsightly as the result that everything is morally good *period*.

Now turn attention to variable C. Its importance (noted earlier) is brought out by imagining how things would stand were it absent. Suppose we restrict our attention to a particular observer, Mary. Does anything at all have the disposition of producing approval in Mary *period*? I should say not. Certainly some things have produced approval in Mary in the past, and certain things can be reliably expected to produce approval in her again. If we consider Mary encountering or reflecting upon certain things (say, acts of generosity), then we might be justified in supposing that she will feel approval. However, in making such observations we will inevitably be including an understanding of Mary's *circumstances*—if only a tacit and vague presupposition of "ordinary circumstances." The fact that we can easily imagine circumstances in which Mary might encounter an act of generosity without feeling approval—because she's being chased by a tiger, for example—demonstrates that in order to specify the dispositional property some restriction must be placed on the circumstances of stimulus, on pain of no dispositional property being picked out.

Prinz's Relativistic Sensibility Theory

In his recent defense of relativistic sensibility theory, Jesse Prinz identifies moral properties as "powers to cause emotions in us" (2007: 89). The "us," it turns out, is left open: After toying with "normal observers" and observers who have "knowledge of relevant facts, and are not under emotional or cognitive influences that are not relevant to the

feature in question into the circumstances: speaking of the disposition to produce R in S in *circumstances that provide full information*. We even might speak of the disposition to produce *fully informed R responses* in Ss in C. On many occasions, such differences in how the disposition is described are of no ontological significance.

case at hand" (2007: 91), Prinz opts to drop all restrictions. He is unfazed by the fact that X may prompt one emotion in one observer and another emotion in another observer; Prinz simply embraces the relativistic view that X may be morally good relative to some observers and bad relative to others.

What of *circumstances* in which a moral property causes an emotion in an observer? Given the lack of restriction on "observer," we cannot make use of the aforementioned solution of considering observers in "ordinary" circumstances, for what are the "ordinary circumstances" of *observers*? One might respond that "ordinary circumstances" can themselves be relativized to types of observer: When the observer in question is a Martian, we mean those circumstances that are ordinary to Martians; when the observer is a Cro-Magnon we mean what is ordinary for them; and when the observer is Mary we mean yet other sets of ordinary circumstances. Not least among the glaring problems with this response is that, for any given observer, there is no fact about the level of generality at which these categories should be drawn. When the observer in question is Mary being chased by a tiger, say, do we look to the "ordinary circumstances" of *a human observer*, or those of *a human observer being chased by a tiger*, or those of *a running human*, or those of *Mary when frightened*, or what?

Seemingly, the only restriction that Prinz places on circumstances is that the observer must be "in good epistemic conditions" (2007: 102). But if our worry is that in failing to specify circumstances Prinz has provided a description of the disposition that is incomplete to such an extent that it fails to denote any property at all, then this slight narrowing of the space of possibilities is unassuaging. X may cause Fred in good epistemic conditions *on Monday* to feel approval, while X may cause Fred in equally good epistemic conditions *on Tuesday* to feel disapproval. Thus the question "Does X cause Fred, when in good epistemic conditions, to feel approval?" has no answer; one must appeal to the inquirer for a more precise question.

Prinz is content to leave circumstances unspecified because (it turns out as his theory develops) the relevant psychological response is not an *emotion*, but a *sentiment*—where "sentiment" is a term of art denoting a dispositional state: the disposition to have an emotion. On a given occasion Mary might lack the emotion of anger, while still having the sentiment of anger. If the relevant sentiment is properly defined, including circumstances of stimulus, then Prinz might not need to specify circumstances in the broader description of the disposition. In other words, the seemingly incomplete description "the disposition to produce response R in subjects of type S" might pass muster if it turns out that "R" surreptitiously specifies circumstances—for example, that "R" is defined as something of the format "the disposition to have emotion E *in circumstances C*."

Let me address some potential puzzlement about this before making my principal criticism. The puzzlement arises because we now seemingly have two dispositions in play: The moral property is a disposition (a "power") and the relevant observer's response is a disposition (a "sentiment"). Whenever we have a response dependency theory, we always have options about whether to discuss dispositions in the world or in

the individual. A dispositional view of color, for example, might claim that redness is the disposition to produce red-sensations in ordinary viewers under optimal viewing conditions. Alternatively, one might say that, for any *x*, *x* is red if and only if ordinary viewers have the disposition to experience red-sensations when observing *x* in optimal viewing conditions. Nobody need deny that both dispositions simultaneously exist. John Heil usefully observes that any "manifestation of a disposition is a manifestation of reciprocal dispositional partner . . . A salt crystal manifests its disposition to dissolve in water by dissolving in water. But this manifestation is a manifestation of both the salt crystal's disposition to dissolve in water *and* the water's reciprocal disposition to dissolve salt" (Heil 2005: 350). In the redness case, all parties can agree to there being dispositions both in the world and in the subject; the pertinent dispute is over whether to identify the former disposition as the property of redness. In the moral case, despite Prinz's tendency to focus on the internal disposition (the sentiment), it is clear that he also wants to identify moral properties with dispositions (though he seems to like the old word "powers") (see Prinz 2007: 89, 92, 107).

Let us return to the matter of specifying circumstances of stimulus, the need for which Prinz thinks he can bypass by making the psychological response itself a disposition (2007: 91–2). But this strategy successfully avoids the charge of incompleteness only if the description of the sentimental disposition is itself complete, and unfortunately the problem just reiterates here, for in his discussion of sentiments Prinz says hardly anything about the circumstances relevant to sentiment dispositions. He mentions that fear of flying is something that manifests itself only when on a plane (2007: 85), but when he comes to the moral sentiments, the need to specify circumstances (if only roughly) seems to have been overlooked. He characterizes resentment, for example, simply as the disposition to feel "bitterness, anger, or contempt" (2007: 86). But does *anyone* have the disposition to feel the occurrent emotion of bitterness *period*? Do you? The natural question is "At what?" But even if "At what?" could be answered—suppose it's specified that we're asking whether you have the disposition to feel bitterness toward ex-lovers—the next question is "In what circumstances?" You might feel occurrent bitterness toward ex-lovers in certain circumstances, but not in other circumstances. (It would be a sad fate indeed if you felt bitterness toward ex-lovers under *any* circumstances.[5])

For reasons that should now be familiar, Prinz's description of the sentiment of resentment, as it stands, fails to denote any property at all in any possible world, and

[5] Prinz also talks of a person's sentiment remaining steady despite the tendency for it to manifest in an occurrent emotion diminishing (2007: 97–8). Frequent exposure to homeless people, for example, may reduce the frequency or intensity of our sympathetic emotions, while our disposition to feel such emotions toward the homeless remains intact. But this in itself reveals a problematic disregard for the role that conditions of stimulus play in defining the disposition in question. If at time *t* passing a homeless person produces strong sympathy in Mary, whereas at $t + 1$ the same stimulus condition does not produce that emotion (but rather a somewhat more extreme exposure is needed in order to prompt Mary's sympathetic emotions), then the dispositional property Mary instantiates at *t* is not the same dispositional property as she instantiates at $t + 1$, and thus the sentiment has not remained steady.

thus, if we take the definition at face value, he has offered an error theory of resentment. And since he has tied moral properties to these sentiments, then if we take *that* definition at face value, he has also offered a moral error theory. My objection here is not simply that Prinz has left his description of the disposition somewhat vague and open-ended. If that were the problem, then virtually every reference to a disposition ever made would be at fault. The problem is, rather, that the description is incomplete in a striking manner that leaves me (at least) with *no idea* how it should be finished, and thus I do not feel inclined to grant the benefit of the doubt that Prinz's description picks out (even vaguely and open-endedly) any property whatsoever.

I confess that I don't really expect this accusation to stick; charging Prinz with unwittingly offering an error theory is really just a cheeky way of pointing out some significant gaps in his metaethical account. Nevertheless, it should lead us to wonder whether accidental error theorists might appear elsewhere on the metaethical landscape.

Firth's Ideal Observer Theory

When a theory of moral dispositionalism offers a description of the relevant disposition that is incomplete to such an extent that we can know without further investigation that there is no such property, I will call this *negligent* dispositionalism. By contrast, *optimistic* dispositionalism is when the description of the disposition leaves it open whether it denotes any actually instantiated property; the advocate of the theory assumes or hopes that it does, but there are serious grounds for doubt. I classify Roderick Firth's *ideal observer theory* as an example of optimistic dispositionalism.

Firth identifies moral goodness with the disposition to prompt approval in the ideal observer, who in turn is defined as omniscient, omnipercipient, disinterested, dispassionate, consistent, and in other respects normal (Firth 1952, 1955).[6] The question with which we are concerned is whether there is *anything* that has this disposition.

The term "the ideal observer" is intended to refer to neither an *actual* individual nor a possible *individual*. Insofar as the characteristics provided are sufficient to locate anyone in modal space, they will presumably locate a number of individuals. Thus the phrase "the ideal observer" is less like "the president of the USA" and more like "the blue whale." When we say "The blue whale lives in the Southern Ocean," we are not referring to an individual whale, but to a kind. This introduces at least a touch of oddity to Firth's theory, for we are supposed to take a token action and consider the response of a *kind* of individual to that action. By analogy, suppose we pointed to a particular school of krill and wondered whether *the blue whale* would have reaction R to that token. What

[6] Firth doesn't actually set out to define *moral goodness* in particular, but rather refers generally to "any moral predicate." He also postpones specifying the relevant kind of reaction, preferring to speak of the ideal observer's "ethically-relevant reaction." I use the term "approval" for brevity.

would that mean? Any blue whale? Some blue whales? Most blue whales? A typical blue whale?

I don't think that there is a settled answer to these kinds of question; it varies with conversational context. When we say "The blue whale is the largest animal ever to live," we don't mean *any* blue whale. The existence of a stunted blue whale considerably smaller than, say, an average sperm whale would not prompt us to retract the statement. Nor must we mean *some* blue whales. If our stunted blue whale individual never surpassed 40 feet long, we would not on that account claim "The blue whale does not surpass 40 feet in length." Nor does it seem correct that we must always mean the *typical* blue whale. If marine biologists were to observe an exceptional whale stay submerged for over an hour, then, even if they were aware that they had witnessed a unique record-breaking event—something that no other blue whale could accomplish—they would not hesitate to claim subsequently "The blue whale can stay submerged for over an hour."

In the case of the ideal observer, the problem posed by analogous questions (some?—all?—most?) would recede considerably if there is a convergence in the ideal observers' relevant responses. But this is exactly the point at which I would like to place pressure on Firth's theory.

We may have some justified beliefs about the effect upon our attitudes of having less information versus having more information, of being calm versus being emotionally aroused, of being selfish versus being generous, and so forth, but we really have *no idea* what a creature would be like with the ideal observer's extreme characteristics.[7] For all we know, complete disinterestedness might lead to the coldest kind of consequentialist calculations, whereby appalling sacrifices will be countenanced for the greater good. Maybe a spot of genocide really would work out for the best *eventually*, and perhaps it is precisely the observer's "idealized" psychology that liberates him or her from those emotions that usually cause us to turn away from that possibility appalled. Or perhaps the ideal observer would be indifferent to the "greater good"; perhaps he or she would be confused by the very idea.

We should also be wary of a lurking fallacy of assuming that because all instances of moral disagreement that we have ever encountered have been due to a deficiency of X (such as true information) among interlocutors, providing X "to an extreme degree" (Firth 1952: 321) will lead to convergence. This is like saying that because a death was caused by a lack of oxygen in the room, death would have been avoided had the room been filled with 100 percent oxygen.

The point in which I am interested is not so much that all ideal observers might turn out to be monsters by our standards, but that the characteristics provided by Firth are insufficient to determine any particular pattern of attitudinal responses. Just as his list

[7] Firth's addendum that the ideal observer is "otherwise normal" seems of little use here. It brings to mind someone describing a divine being as all-seeing, all-knowing, all-powerful, infinitely loving, eternally existing, the creator of the universe … but *otherwise just like you and me*.

is too general to pick out an individual in modal space but rather picks out a kind, so too it may be too general to pick out a kind with a determinate pattern of (dis)approval but rather picks out a kind with attitudinal variation in this respect. There may be no fact of the matter about what an ideal observer would approve of, any more than there is a fact of the matter about whether the ideal observer prefers vanilla to chocolate ice cream. Consider an actual token action ϕ. If we examine the closest possible worlds at which there are ideal observers, then perhaps some of them disapprove of ϕ, some are neutral, and some even approve (even when in the same circumstances). If this is so, then does ϕ instantiate the dispositional property of prompting approval in the ideal observer? If we mean *all* ideal observers, then the answer is "no."[8] It may be that *some* actions are like ϕ in this respect or it may be that *all* actions are. Despite its strength, I do not find the last possibility absurd—it does not seem implausible that there exists *nothing* about which equally ideal Firthian observers will agree—in which case *nothing* will be morally good or bad.

Firth in fact admits to one of the premises of this argument. When pressed upon the question of convergence by Richard Brandt (Brandt 1955: 408–9), Firth admits that if there could be two ideal observers with different or opposed reactions to an act, "it would follow . . . that the act in question would be neither right nor wrong" (Firth 1955: 415). Firth rejects the antecedent, however, by claiming that divergent attitudinal responses imply differences in the traits used to identify the ideal observers.

But Firth is far from convincing on this point. The traits that he provides for the ideal observers clearly aren't sufficient to ensure a convergence on favorite ice cream flavor; perhaps they aren't sufficient to ensure a convergence on attitudes of approval and disapproval either. The question of what shared psychological traits are sufficient to ensure a convergence in (dis)approval is to a large extent an empirical matter, many of the details of which remain unknown. There is a growing body of literature revealing that the things that can influence an individual's morally relevant attitudes can be quite surprising. We might not have supposed, for example, that a person's tendency to act dishonestly can be enhanced by her wearing sunglasses or being placed in a dimly lit room (Zhong et al. 2010). Nor might we have guessed the effect of hand-washing on a person's moral evaluations (Schnall, Benton, et al. 2008). We might not have appreciated how easy it is to manipulate someone's moral opinions by placing him in a messy environment—for example, in the presence of a dirty tissue (Schnall, Haidt, et al. 2008).

Firth thinks that the characteristics he uses to pick out the ideal observer in modal space are sufficient to (a) ensure convergence and (b) get intuitively correct results (for

[8] For further discussion of this point, see Carson (1984, 1989) and objections by Taliaferro (1988). If, alternatively, we interpret the question as asking whether the action would prompt approval in *some* ideal observers, then the answer is presumably "yes." The problem with this, however, is that it is not unreasonable to suspect that for just about anything there is *some* ideal observer that approves of it. This would not make the proponent of ideal observer theory an unwitting error theorist, but would nevertheless be a kind of reductio: of implying that just about everything is morally good. Relativizing one's ideal observer theory (see Carson 1984, 1989) is one obvious response, though such a reaction brings its own set of problems.

example, the ideal observers do not turn out to all be Nazi sympathizers); he argues that all instances of moral disagreement which he has observed or which he can imagine are the result of differences in belief, or selfish interests, or self-referential emotions. Let us consider the plausibility of this in light of one of the empirical cases just cited: Suppose that there are two people making moral judgments about ϕ, and one of them is in the presence of a disgust-prompting dirty tissue which influences him to judge ϕ negatively, while the other is not. Certainly the two have different beliefs (one thinks "There's a dirty tissue," while the other does not think this), but this fact doesn't mean that they cannot both be equally ideal by Firth's lights. As far as beliefs go, Firth says that the ideal observer must be "omniscient with respect to the non-ethical facts"; but this is not to say that all ideal observers must have *the same* beliefs. We should presume that both ideal observes will be well-informed about experimental psychology: They will know about all of the aforementioned studies, including the "Schnall, Haidt, et al. 2008" paper which demonstrates the influence a dirty tissue may have on a subject's moral attitudes. Moreover, we can assume that the disgusted person realizes that he is being manipulated in just this way. But does this knowledge make his disgust (and connected moral assessment) dissipate? Perhaps; perhaps not. Attitudes prompted manipulatively in the setting of a psychology lab often survive the debriefing session (Ross et al. 1975; Nisbett and Ross 1980). Even medical placebos sometimes work in conditions of full information (Park and Covi 1965; Aulas and Rosner 2003). It is not my intention to persuade anyone that the disgust *would* remain in situations of full-information, only that it is an empirical matter against which Firth would be imprudent to bet the farm.

If neither of our two imaginary persons must be suffering from doxastic failing, Firth may instead claim that the subject whose moral assessment is influenced by the dirty tissue has become less dispassionate or less disinterested. This can seem plausible if one equates dispassionateness with lack of emotion, for it seems highly likely that the dirty tissue influences moral assessment only via arousing the emotion of disgust. However, such an equation would be a mistake. Firth's notion of dispassionateness pertains to the absence of "*particular* emotions," which are defined as emotions that are "directed toward an object only because the object is thought to have one or more essentially particular properties" (1952: 340)—where "particular properties" are those "which cannot be defined without the use of proper names" (1952: 338).[9] I see no ground for assuming that the disgust one might feel at the presence of a dirty tissue must take the form of a particular emotion; one's emotion might be directed at dirty tissues—or, more likely, the associated bodily fluids—in general.

The more general point to which I should like to draw attention is the fact that at present nobody knows too much about the psychological mechanisms through which these kinds of subtle influences on morally relevant attitudes work, and it would be a

[9] For convenience, Firth includes pronouns such as "I," "here," and "this" as proper names.

hasty empirical bet to assume that differences in attitude must always entail a difference in the traits of the Firthian ideal observer. So on the question of whether he is committing himself to an error theory, Firth becomes a hostage to empirical fortune.[10]

Scanlon's Hypothetical Contractualism

Contemporary versions of hypothetical contractualism face analogous challenges. Here, it is not the attitude of a (hypothetical) kind of individual that counts, but rather the collective response of a (hypothetical) group of persons. Many objections to such views have been voiced in the literature, but it is seldom appreciated that the dispositional description proposed by contractualism may simply fail to denote any property at all.

Thomas Scanlon, for example, writes that "an act is wrong if its performance under the circumstances would be disallowed by any set of principles for the general regulation of behavior that no one could reasonably reject as a basis for informed, unforced general agreement" (Scanlon 1998: 153). Unlike Firth, Scanlon at least wears his quantifiers on his sleeve: Instead of speaking somewhat mysteriously of "*the* set of principles that would be adopted by *the* group," he makes clear that he means "*any* set of principles that could not be reasonably rejected by *any* group." But even so, is there *any* action that would be disallowed by any such set of principles? The question, again, is one of convergence.

Suppose, for the sake of introducing the argument, we were just talking about sets of principles that would be accepted or rejected by possible *groups of humans*, with no further qualification imposed on what kind of humans; we are, in other words, including Vlad the Impaler and his henchmen, kamikaze pilots, drunken Vikings, suicidal nihilists, the woefully stupid, the willfully annoying,[11] and so forth. Presumably, the "sets of principles for the general regulation of behavior" that these human groups might endorse will vary wildly and may not bear much resemblance to those sets that will tempt civilized folk. Is there *any* action that would be disallowed by *any* of these sets of principles? I see no grounds that should incline one to answer in the positive. But if this is correct, then, if by "... is morally wrong" we mean "... is such that its performance under the circumstances would be disallowed by any set of principles for the

[10] Alternatively, Firth could try to rule out these kinds of influence on ideal observers' attitudes by specifying the hypothetical *circumstances* of the ideal observers' judgment in a way that excludes such possibilities. In other words, he could say that X is morally right iff the ideal observer would feel approval toward X in circumstances where there are no dirty tissues nearby, where he is in well-lit conditions, where he has not been prevented from washing his hands, and so on. It is, however, difficult to see how that "... and so on" is going to be cashed out. Moreover, it is worth noting that one of the curious things about Firth's dispositionalist account is that the *conditions* of stimulus are never mentioned. It is as if he thinks that they just don't matter at all. But of course they do; and one thing I've already argued is that any dispositionalist who neglects to specify conditions of stimulus (at least a ballpark estimate) is on the fast track to an error theory.

[11] Among the ranks of the willfully annoying I include those imaginary humans who choose principles of behavior on the sole basis of refuting popular metaethical theories.

general regulation of behavior that no group of humans would reject," there exists nothing that is morally wrong.

Clearly, then, all that stands between Scanlon and the error theory is the all-important restriction he places upon the type of sets of principles admitted: those that no group "could reasonably reject as a basis for informed, unforced general agreement." And our attention should immediately be drawn to the word "reasonably." The natural supposition is that it is this qualification that acts as the principal bulwark against a slide to moral skepticism, but along with this supposition comes the suspicion that Scanlon cannot simply help himself to the notion in advance of having refuted that very skepticism. If what is "reasonable" in this context implies a substantive moral framework, then Scanlon clearly begs the question.[12] Perhaps to our taxonomy of negligent dispositionalists and optimistic dispositionalists we should add question-begging dispositionalists. But it isn't begging the question of which I wish principally to accuse Scanlon, but rather an ungrounded optimism. Or, perhaps speaking more carefully, I suspect him of flitting between begging the question and ungrounded optimism without finding a stable point between.

Scanlon considers the following imaginary case:

> Suppose that Jones has suffered an accident in the transmitter room of a television station. Electrical equipment has fallen on his arm, and we cannot rescue him without turning off the transmitter for 15 minutes. A World Cup match is in progress, watched by many people, and it will not be over for an hour. Jones's injury will not get any worse if we wait, but his hand has been mashed and he is receiving extremely painful electrical shocks. Should we rescue him now or wait until the match is over? Does the right thing to do depend on how many people are watching—whether it is one million or five million or a hundred million? (Scanlon 1998: 235)

Scanlon goes on: "It seems to me that we should not wait, no matter how many viewers there are, and I believe that contractualism can account for this judgment" (1998: 235). He backs up this opinion by appeal to the *unfairness* of imposing such a sacrifice upon Jones in order to avoid the proportionally lesser inconvenience to each individual viewer. However, in order to establish that leaving Jones to suffer is morally wrong (by Scanlon's own lights), it is insufficient to show that endorsing a set of principles that disallows the imposition of this suffering is a reasonable choice. (I myself feel confident that it is reasonable, as does Scanlon, as, presumably, do most readers.) Rather, it must be shown that any group of persons (aiming at informed, unforced general agreement) that rejected any such set of principles *would be unreasonable*. This latter conclusion evidently doesn't follow from the former; from the fact that someone could reasonably accept X it hardly follows that anyone who rejects X is unreasonable. But Scanlon has nothing else to say to convince us of the crucial proposition, and it does not seem difficult to generate doubt about it.

[12] "It would clearly render [Scanlon's] position uselessly circular if the fact that a putative principle permitted agents to act wrongly were to be adduced as a reasonable ground for rejecting it; for the procedure is supposed to help us identify what courses of action are wrong" (Baldwin 2002: 99).

Empirical evidence reveals a great deal of cross-cultural variation in norms of *fairness* (see Henrich et al. 2004), and it is entirely possible that we could locate actual normative frameworks that will reject principles prohibiting us from waiting till the match is over before rescuing Jones. I will, however, stick with an imaginary case in order to make the point crisply. Suppose a group of persons (let's call them "Stoic sages") believe in some ubiquitous providential divine plan—in such a way that they consider everyday "accidents" to serve some great (though mysterious) purpose, interference with which is to be avoided. Holding this belief, centrally and sincerely, does not obviously exclude the possibilities that these Stoics may seek practical principles and may value informed general agreement. (After all, the real Stoics did maintain distinct political views.) When faced with the unfortunate trapped Jones, the Stoics feel disinclined to step in to upset the unfolding of the divine plan; in fact, they steadfastly reject any set of principles that demands or even permits interference. Is the Stoics' rejection of any such set of principles *unreasonable*?

In a sense, of course it is. The problem, however, is that this sense is one that is ineliminably informed by our own substantive moral beliefs and principles, and thus is not one to which Scanlon can appeal, on pain of endorsing a question-begging dispositionalism. But the other horn of the dilemma is no less damaging: If Scanlon employs a thinner and less morally loaded notion of *reasonableness*, then there seem no grounds to exclude the Stoics' choice as unreasonable. My firm suspicion is that we can pull this move over and over again, for *any* action one cares to mention: We can always locate a hypothetical group of persons with sufficiently wacky beliefs about how the universe works, or harboring sufficiently aberrant desires, or committed to sufficiently bizarre values, or inhabiting sufficiently atypical circumstances, that they will be willing to reject any set of principles we care to imagine, *without our being able to make the charge of unreasonableness stick*. If this is so, then there exists *no* set of principles for the general regulation of behavior that *no one* could reasonably reject as a basis for informed, unforced general agreement.

It is possible that Scanlon, and hypothetical contractualists in general, might yet find some kind of plausible rejoinder to this line of objection. My point is that they are yet to do so, and so at this stage we must consider the conviction that the requisite notion of *reasonableness* is forthcoming to be another instance of optimism being the only thing standing between endorsement of the theory and the moral error theory.[13]

[13] I am also inclined to accuse Michael Smith (1994) of supporting an optimistic dispositionalism. Smith argues that S has a normative reason to ϕ iff a fully rational counterpart of S would advise S to ϕ. He then argues that ϕing is morally right only if *everyone's* fully rational counterpart would come down on the side of ϕing. The latter is a conceptual claim, leaving open the substantive question of whether everyone's normative reasons do in fact converge in the necessary manner. Smith is here consigned to a footnote because I have criticized him on this point before (Joyce 2001: 88–95; see also Sobel 1999), and in any case what I have said against Firth and Scanlon gives a pretty good hint of what I will say again against Smith. What is distinctive about Smith in the present context is that, if faced with the failure of the convergence premise, he seems willing to embrace the error theory. (See Smith 1994: 187–9, 2002, 2006.) Doubts about convergence also lead me to regard Frank Jackson (1998) as a potential accidental error theorist. See Robinson (2009) for criticism of Jackson's presuppositions about convergence.

Virtue Ethics

The virtue ethicist is another prominent example of the kind of concessive opponent of the moral error theorist that has been under discussion. The virtue ethicist may share many or all of Mackie's misgivings about objective values and objective prescriptions. "Certainly," she may concede, "if morality were like *that* then we should all be error theorists; but morality is not like that, or, at least, need not be like that." And the virtue ethicist will then point to the Ancient Greeks—and Aristotle in particular—as providing an exemplar of an ethical system that got along very nicely without all those problematic objective values and prescriptions ruining everything. (The *locus classicus* of this view is Anscombe 1958.) Unlike deontological ethical theories that begin with the action-oriented question "How ought one to act?," or teleological ethical theories that begin with the value-oriented question "What is of intrinsic value?," virtue ethics begins with agent-oriented questions like "What kind of person should one be?" or "What is the good life for a human being?"[14] It is thought that this fundamental difference in starting point promises to immunize virtue ethics from the kinds of error-theoretic worries that dog deontological and teleological theories. But is this true? Does virtue ethics really represent a smooth escape route from the threat of a moral error theory?

It suffices for an answer in the negative if serious doubt arises as to whether there even *are* any of the entities that the virtue ethicist refers to as "virtues." If there simply are not any such things, then all the virtue ethicist's distinctive assertions—such as "The virtue of honesty is an important part of human flourishing" or "Albert Schweitzer was more virtuous than Albert Speer"—and all deontological and teleological talk that virtue ethicists allow as derivative upon virtue talk—will fail to be true. In other words, if the virtue ethicist bases his or her theory on the claim that

$(\forall x)x$ is a virtue iff x is a P,

but it turns out that nothing satisfies the predicate "… is a P," then the virtue ethicist is in fact proposing an error theory.

One possible way that this might happen is if the virtue ethicist's conception of a virtue ineliminably presupposes a badly flawed image of human psychology. Gilbert Harman and John Doris have both argued that the existence of the kind of entrenched personality traits upon which virtue ethics depends is cast into doubt by empirical evidence in support of "situationism" in social psychology (Harman 1999, 2000; Doris 2002). This is a controversial claim (see Merritt 2002; Sreenivasan 2002), and it is not a strategy that will be further explored here.

Rather, I have doubts about the virtue ethicist's starting presuppositions: the eudaimonia-oriented questions—*What kind of person should one be? What is the good*

[14] Deontological theories put duty first, and define value in relation to duty. Teleological theories (such as utilitarianism) put value first, and define duty in relation to value. See Broad (1930: 277ff).

life for a human being?—which are supposed to provide such a trouble-free point of departure when compared to the rival deontologists' and teleologists' guiding inquiries. My understanding is that when we direct these questions at traits of character (asking, for example, *What kind of character traits must one cultivate in order to be the kind of person one should be?*), then the virtues have the theoretical role of *answers* to these questions. But my contention is that it is entirely possible that these questions have no answers and thus there are no virtues (as conceived by the virtue ethicist).

This concern can be brought into focus by parodying the virtue ethicist's questions: *What kind of ice cream flavor must one prefer in order to be the kind of person one should be? What ice cream preference contributes to the good life for a human being?* Let us assume, not unreasonably, that it is acceptable to choose ice cream flavors on the basis of gustatory whim. It is *possible* that certain flavors are better for one's health, or better for the environment (if they use sustainable ingredients, say), or better for the wider community (if their production eschews exploitative practices, say)—but let's assume that all such potential complications come to naught and that one can select on the basis of taste alone. Then I would know what flavor *I* should prefer,[15] but there would be no flavor that "one" should prefer, and no flavor that one must prefer in order to be the kind of person that one should be. To equate the predicate "…is P" with "…is the ice cream flavor that one must prefer in order to be the kind of person that one should be" would be to endorse (perhaps unwittingly) an error theory about P-discourse.

In order for the virtue ethicist's questions to fare better, the kind of life that "one should live" cannot be similarly a matter of whimsical choice and cannot change from individual to individual; it must be grounded in something shared by all humans. This is a problem, since there are many images of the good life: the life of the Buddhist monk, of the hedonistic consumer, of the intellectual, of the Stoic sage, of the noble savage, and so on. At this point, the Aristotelian virtue ethicist will often appeal to *human nature* in order to privilege one kind of "good life" that is shared by all. The virtue ethicist Rosalind Hursthouse writes: "A virtue is a character trait that human beings, given their physical and psychological nature, need to flourish (or to do and fare well)" (1995: 68). Clarity demands that we distinguish the kind of flourishing that (supposedly) derives from human nature from any alternative and competing visions of human flourishing that an individual or group might (or might not) embrace, and since the pertinent difference here appears to be that some visions of flourishing might be *chosen* while the one derived from human nature is bestowed upon us whether we like it or not, I shall refer (somewhat clunkily) to the nature-given account of the good life as "non-chosen human flourishing." This kind of human flourishing should be no more troubling (the virtue ethicist avows) to our naturalistically inclined philosophical temperaments than the notions of *antelope flourishing* or *petunia flourishing*, which can be derived from accounts of antelope nature and petunia nature, respectively. "A correct conception of the virtues must be at least partly shaped by a correct conception

[15] Ben & Jerry's Rainforest Crunch—now, alas, discontinued. (See Ben & Jerry's "Flavor Graveyard.")

of healthy growth and development which in part constitute our flourishing" (Swanton 2003: 60). The virtue ethicist will, moreover, stress the *social* nature of our species, in the expectation that the more prosocial virtues, like generosity and friendship, will be contributors to non-chosen human flourishing. "We are naturally sociable creatures who like to have friends and want to be loved by friends and family" (Hursthouse 1987: 226).

The core of my skepticism about this is that there remains abundant room for reasonable doubt that the facts of "human nature" are going to play out in the determinate way that the virtue ethicist assumes. One can allow (if only for the sake of argument) that it is legitimate to speak of "human nature" and hence "human flourishing," but nevertheless humans are the most psychologically plastic organisms we have ever encountered, and thus the "end" of human flourishing may provide only a minimal constraint on lifestyle decisions, and no constraint at all on character traits. Humans are without doubt obligatorily gregarious organisms, and so one might reasonably claim that living in some sort of community of fellows is an "end" that has been conferred upon humans by nature. But what degree of specificity of character traits is determined by this "end"? Hitler had loyal and sincere admirers; Genghis Khan was surrounded by good mates; perhaps even Jack the Ripper was a solid family man. The idea that the sociality inherent in human nature cannot be satisfied in a restricted domain, while coupled with cold disregard and astounding cruelty toward anyone lying outside the favored sphere, strikes me as a romantic misapprehension. To put the point provocatively: It is not foolish to declare that Hitler's character traits were just as true to his nature as a social organism as Mother Teresa's.[16] (Let us not forget that the Nazi war machine required an enormous amount of interpersonal cooperation, much of which was motivated by strong prosocial feelings.) If this is so, then there may be no specific set of character traits that is underwritten by our social nature.

One might be tempted to respond that there is surely *something* in common between Hitler and Mother Teresa with respect to their social skills: some very general and minimal interpersonal faculties operative in any human who manages to have any kind of successful relationship with his or her fellows. Maybe, then, these very minimal traits might count among the virtues? But this is hardly a line that the virtue ethicist will find attractive, for if even Hitler and Jack the Ripper turn out to have the social virtues, then we've surely seriously lost track of the point of endorsing virtue ethics. Second, it is far from obvious that the kinds of minimal social skills manifest by anyone capable of maintaining any sort of meaningful interpersonal relationship are going to count as *character traits* in the requisite manner. The Aristotelian virtue ethicist will usually embrace a view of character traits according to which they are "relatively long-term stable dispositions to act in distinctive ways . . . involving [*inter alia*] habits of desiring" (Harman 1999: 317). Even if humans are by nature social organisms, and thus need

[16] Though it *would* be foolish to think that my saying this indicates any glimmer of tolerance toward Hitler.

certain traits in order to flourish as social organisms, it doesn't follow that there is a set of *character traits* (in the sense just described) that humans need in order to flourish as social organisms—any more than there is an ice cream preference we need in order to flourish as social organisms. Thus, if by *virtue* the virtue ethicist specifically means "a character trait needed for non-chosen human flourishing," then we are once more looking at a potential error theory.

None of this is to deny that we are free to create and embrace more substantial visions of flourishing and the good life, for many of which the cultivation of specific character traits will certainly be necessary (either causally or constitutively). For example, one might maintain that the good life consists of living like a Buddhist monk—a kind of life for which (a) the claim that it is "natural" for humans is highly implausible, and (b) certain character traits, like irenic acceptance, are necessary. The problem, though, is that one individual's or community's robust vision of the good life will differ from another's, and the character traits needed to succeed at one life may diverge from those needed to succeed at another. It may well be true that for any person, x, there exists a good life, y, and exists a certain set of character traits, z, such that z is necessary for y. But we must be careful not to commit a quantifier-shift fallacy of flipping this round and thinking that there exists a set of character traits, z, and exists a good life, y, such that for any person, x, y is x's good life and z is necessary for y.

We have seen that so long as *a virtue* is defined as "a character trait necessary for non-chosen human flourishing," there may be no such thing. I haven't tried to establish that there is no such thing, but merely to expose the presence of reasonable doubt. The virtue ethicist may attempt to alleviate this doubt by weakening the definition to "a character trait that *tends to* contribute to non-chosen human flourishing." ("[T]he claim is not that being virtuous guarantees that one will flourish . . . Virtue is only a reliable bet; it will probably bring flourishing" (Hursthouse 1987: 230).) But this does not necessarily help. After all (sticking with a parody that no doubt grows repetitive but remains instructive): There is no ice cream flavor preference that even *tends* to contribute to non-chosen human flourishing. So why assume that there is any such set of character traits?

Rather than answer that question directly, let me try another tack. Suppose that there *are* some character traits that tend to contribute to non-chosen human flourishing. I noted earlier the multiplicity of alternative visions of the good life, and we can suppose that there are also sets of character traits that tend to contribute to these alternative images. One version of the good life encourages friendliness, say, while another version (seemingly as legitimate as the first) urges remaining aloof; one encourages turning the other cheek, while another urges vengeance, and so forth. Will the virtue ethicist be satisfied merely to persuade us of the terminological stipulation that one among these competing sets—the one that is "nature-given"—can be given the label "the moral virtues," and leave it at that? I should think not. The moral virtues are expected to carry some practical weight, some normative force, some extra authority. Suppose this is considered to be not merely a contingent feature of virtue, but is taken

to be an additional essential quality; a virtue is defined not merely as a character trait that probabilifies non-chosen human flourishing, but a trait whose cultivation carries more normative weight than any other character traits which probabilify any alternative chosen visions of flourishing.

However, I again find myself hesitant to believe that there exist *any* character traits that enjoy this attribute. Indeed, if any kind of substantive normativity is made a defining feature of the virtues then this may lead straight to an error theory, for it is hard to see how non-chosen human flourishing will supply those traits that conduce to its satisfaction with any normative relevance whatsoever.

Let us think of this in terms of proper functions. The proper function of a wire coat hanger is to hang up clothes, which gives license to a variety of normative-sounding language, such as "A coat hanger ought to support clothes" and "A good coat hanger supports clothes well." This may provide one with various reasons *if one wants to hang up clothes*, but if one has no such interest, but rather has an interest in, say, retrieving one's dropped keys from the drain, then using the coat hanger as a fishing hook (destroying it in the process) is entirely legitimate. This would not be an instance of there being two competing reason-conferring functions—one of which outweighs the other. Rather, given one's interests, the proper function of the coat hanger—what the hanger is "supposed to do"—carries no weight in one's deliberations whatsoever. There is not a slightest drop of true normativity (independent of the agent's antecedent interests) that can be squeezed from the proper function of the coat hanger.

Suppose that non-chosen human flourishing really does require the cultivation of certain prosocial character traits. But there will also be a range of alternative ends that a person might genuinely prefer (say, the life of selfish hedonism, or the life of an ascetic hermit, or the life of a reclusive intellectual) which require the cultivation of different sets of character traits. The question is not merely why the former end must trump any of the latter, but why the former end, in and of itself, constitutes any kind of practical consideration at all. If a hermit withdraws from society—letting his prosocial character traits wither, abandoning the end of nature-bestowed human flourishing in favor of an alternative vision of flourishing—must this constitute *a mistake*, any more than the person who uses a wire coat hanger to retrieve her keys makes a mistake?

Lest it be thought that the example of a hermit is too extreme and unusual to be bothersome, we should remind ourselves that the same point could be made of any lifestyle one cares to think of. We can imagine someone who prefers the end of living by-and-large like an upstanding citizen but with occasional self-serving exceptions (even at serious cost to others) when the chances of incurring punishment are low. Just picture any behavior at all that seems intuitively morally wrong—whether mild or dramatic: We can imagine someone constructing and preferring a lifestyle that occasionally allows this behavior, and cultivating the set of character traits conducive to that end. The question to which I am drawing attention is: Why, in such circumstances, does the "natural" end of human flourishing, from which the preferred end deviates either a little or a lot, furnish any practical authority at all?

One might be inclined to answer that organisms are more likely to achieve well-being and fulfillment if they satisfy the ends laid down by nature (where this is not a trivial claim derived from using "well-being" and "fulfillment" as synonyms of "human flourishing"). But, first, this is an empirically dubious claim. Natural selection may well have forged us as creatures that *strive* to achieve a sense of well-being; but there would be little evolutionary mileage in creating creatures that actually *achieve* a sense of long-standing well-being upon attaining fitness-enhancing goals. The plausibility of the hypothesis that the true road to well-being and fulfillment is to live like a Buddhist monk is in no way undermined by the observation that such a life probably represents a dramatic departure from the kind of "human flourishing" laid down by nature. Second, even if it were true that non-chosen flourishing reliably leads to well-being and fulfillment, if this is the sole ground for recommending the character traits conducive to that end, then it is not the flourishing per se that matters, but the states of well-being, and so on, that reliably accompany it. What, then, would there be to exclude the discovery of an alternative (possibly vicious) means of achieving that same well-being, perhaps more efficiently and abundantly?

I conclude, therefore, that if a virtue is defined as a character trait that is necessary for, or probabilifies, non-chosen human flourishing *and thus has normative weight*, then there are additional grounds for doubting that there exist any such things at all. There may be versions of virtue ethics that do not include such a claim—that are not even based on any claim like "The virtues are those character traits that tend to contribute to human flourishing"—and I freely admit that the objections I have raised here do not apply to any such versions.

Conclusion

My ambitions have been more modest than they might appear. I have discussed several well-known metaethical theories in a critical voice, underlining the places where proponents, if they are not careful, will commit themselves to a moral error theory. It bears repeating that my calling these philosophers "accidental error theorists" is not to be taken too literally; it is really just a slightly mischievous way of drawing attention to a pattern of defect in their theories. This is not so much a damning critique of these types of theories as a plea for greater specificity. I do not mean to suggest that adequate specificity may not be supplied which will allow these theories to avoid the pitfalls I have set before them. It should, however, be underlined that even if moral dispositionalists and virtue ethicists can avoid the charge that they are unwittingly error theorists, it doesn't follow that they thereby defeat the error theorist. Even if, for example, we are satisfied that in the equation "Moral goodness = property P" the right-hand side succeeds in denoting an instantiated property, the error theorist can still object to the adequacy of the equation as a whole. Indeed, this latter kind of argument may well be the strongest strategy for the error theorist to pursue. My intention has been merely to show that some metaethical theories threaten to disintegrate before the need for debating the plausibility of the identity claim even arises.

4

Metaethical Pluralism

How both Moral Naturalism and Moral Skepticism may be Permissible Positions

Introduction: Some Definitions

This essay concerns the relation between two metaethical theses: moral naturalism and moral skepticism. It is important that we distinguish both from a couple of methodological principles with which they might be confused. Let us give the label "Cartesian skepticism" to the method of subjecting to doubt everything for which it is possible to do so—usually by introducing alternative hypotheses that are consistent with all available evidence (such as brains in vats). Let us give the label "global naturalism" to the principle that requires of any item which we admit into our ontology that it "fits" (in some manner or cluster of manners to be specified) with our naturalistic scientific worldview. One might be both a Cartesian skeptic and a global naturalist, if the latter principle is something that has survived the former test procedure. Alternatively, one might have adopted global naturalism for some other reason, while having little patience with the Cartesian method of doubt.

Moral naturalism is the metaethical view that moral entities (for example, properties such as *goodness* and *evil*) fit within the scientific image of the world. The moral naturalist will probably be a global naturalist, but need not be: It is consistent with allowing non-natural entities into one's ontology that one happens to think that moral properties are of the natural variety.

Moral skepticism denies that moral entities fit within our scientific worldview. One way of denying moral naturalism is to be a moral error theorist: to hold that our moral discourse attempts to make reference to moral properties, but these properties do not exist.[1] Another way of denying moral naturalism is to be a noncognitivist: to hold that our moral discourse was never really in the business of referring to moral facts or properties in the first place, and *ipso facto* such facts or properties are not naturalistic.

[1] The usual error-theoretic strategy here is to argue that moral properties have features that no naturalistic property can have (for example, some strong kind of practical authority) and thus the moral naturalist's characterization of moral properties must be inadequate. (See Joyce 2001; Mackie 1977.) Alternatively, the error theorist might accept the moral naturalist's characterization of the moral (for the sake of argument) and argue that there are *still* no such properties. For a study of the latter strategy, see Joyce (2011a).

In this paper, the label "moral skepticism" denotes the disjunction of these two theses. Neither the error theorist nor the noncognitivist must be committed to global naturalism, but usually will be; indeed, this commitment will often be a motivating factor of their metaethical views. The error theorist who is a global naturalist will typically deny the existence of moral properties precisely because (he thinks) these properties fail to fit with a naturalistic worldview. Likewise, those who turn to noncognitivism often do so because they cannot see how moral properties could find a place in a naturalistic order.

(There is a third way of denying moral naturalism which is not a kind of moral skepticism at all. The *moral non-naturalist* denies that moral entities fit within our naturalistic worldview but does not judge this a mark against their existence or our epistemic access to them; moral facts are acceptable non-natural facts. I shall not be discussing this metaethical view in this essay.)

"Skepticism" is a pliable term, and the way I am using it here will not jibe with the way all others use it.[2] I have already stipulated how the term will be employed, but it is probably useful to make a couple of further clarifications. First, I have described both the error theorist and the noncognitivist as *denying* something: that moral entities are to be admitted into our ontology. This stands in contrast to a classical kind of skeptic who neither affirms nor denies but simply doubts. The skeptic I have in mind doubts X in a strong sense of affirming the denial of X (and is thus, in the classical vocabulary, a negative dogmatist). Second, my moral skeptic is not playing the Cartesian game of doubting moral naturalism simply because alternative hypotheses can be conjured. My moral skeptic has probably accepted the naturalistic worldview—indeed, employs it as a premise for skepticism. Thus the skeptic takes himself or herself to have positive grounds for doubting moral naturalism beyond the unimpressive pastime of imagining far-fetched (but unfalsifiable) scenarios in which moral properties are absent. This is a *naturalistic* skepticism, not a Cartesian skepticism.

Moral naturalism and moral skepticism are certainly not contradictories; they do not exhaust the space of metaethical possibilities. (I have already mentioned one alternative: non-naturalism.) But it is usual to think of them as contraries, for surely to embrace one position is to reject the other. In this paper I want to investigate this latter assumption closely. I will explore the possibility of a certain kind of metaethical pluralism, whereby seemingly contrary metaethical positions enjoy equal claims to legitimacy—and not just in the sense that we are not (yet) in a position to know which position is correct, but in the more interesting sense that there is no fact of the matter about which is correct.

[2] It is difficult to choose a label that denotes the disjunction of the error theory and noncognitivism. "Moral anti-realism" would serve in some people's books, but I prefer to reserve this term to cover the error theory, noncognitivism, *and* a kind of moral naturalism according to which moral facts are in some manner constructed by human attitudes (as opposed to being objective features of the world). To my terminological tastes, some moral naturalists are moral realists and some are moral anti-realists. The disjunction of the error theory and noncognitivism is called "skepticism about moral truth" by Walter Sinnott-Armstrong in his taxonomy of moral skepticisms (Sinnott-Armstrong 2006: 11).

Lewis's Pluralism

I take as my point of departure a series of intriguing and characteristically astute comments made by David Lewis at the close of his 1989 paper "Dispositional Theories of Value. Part II." In the course of this paper Lewis has developed and advocated a kind of moral naturalism: a dispositional theory according to which values are those things which we are disposed, under certain idealized conditions, to desire to desire. Worried that this theory makes values contingent—that we feel uneasy about a theory that allows that benevolence (say) is, *but might not have been*, a value—Lewis toys with identifying values with what we are *necessarily* disposed to desire to desire. This amendment would probably soothe the worry, he acknowledges, but it would be at a price: There is probably *nothing* that we are necessarily disposed to desire to desire; thus to identify moral properties with this dispositional property would leave us with a moral error theory.

It is not Lewis's dispositional theory per se that interests me here, but what he goes on to say next. The version with the necessity operator is, he thinks, that which "best captures what it would take for something to perfectly deserve the name 'value'" (Lewis 1989: 136). But this perfect deserver leaves us with skepticism. However, "there are plenty of imperfect deservers of the name" (136).

> Strictly speaking, nothing shall get the name without deserving it perfectly. Strictly speaking . . . genuine values would have to meet an impossible condition, so it is an error to think there are any. Loosely speaking, the name may go to a claimant that deserves it imperfectly. Loosely speaking . . . [t]here are values, lots of them. (136–7)

He goes on:

> What to make of the situation is mainly a matter of temperament. You can bang the drum about how philosophy has uncovered a terrible secret: there are no values! . . . Or you can think it better for public safety to keep quiet and hope people will go on as before. Or you can declare that there are no values, but that nevertheless it is legitimate—and not just expedient—for us to carry on with value-talk, since we can make it all go smoothly if we just give the name of value to claimants that don't quite deserve it . . . Or you can think it an empty question whether there are values: say what you please, speak strictly or loosely. When it comes to deserving a name, there's better and worse but who's to say how good is good enough? Or you can think it clear that the imperfect deservers of the name are good enough, but only just, and say that although there are values we are still terribly wrong about them. Or you can calmly say that value (like simultaneity) is not quite as some of us sometimes thought. Myself, I prefer the calm and conservative responses. But as far as the analysis of value goes, they're all much of a muchness. (137)

Perhaps the moral error theorist has a different temperament from that of Lewis. I confess that I have long been drawn to the error-theoretic version of moral skepticism, and it would be untruthful to claim that this is entirely the product of having been persuaded by solid philosophical argumentation. Maybe it is a mischievous streak in

me; perhaps it is an excitement about striding into uncharted intellectual territory having cast off familiar assumptions. Who knows? I wouldn't attempt to publicly analyze the antecedents of my skeptic-leaning temperament any more than Lewis sought to openly scrutinize his own conservative preferences. The important points are (i) that temperaments opposed to Lewis's exist (exhibit A = myself), and (ii) that Lewis thinks our temperaments determine "what to make of the situation." Thus he makes one thing of the situation—moral naturalism—while I make another—error-theoretic skepticism. Is this a reasonable conclusion? And, if so, then where (if anywhere) do we go from here?

Lewis is certainly not saying that *all* there is to the debate between the moral naturalist and the moral error theorist are their respective temperaments. Both parties must first have some decent arguments on their side. For example, we can assume that no amount of "temperament" favoring the existence of magic will make a naturalistic endorsement of magic reasonable. Likewise, no amount of "temperament" opposed to the existence of big hairy spiders is going to make doubting their existence epistemologically acceptable. From the perspective of global naturalism, some things clearly exist (spiders) and some things clearly do not (magic). Where things get interesting is in the gray area, where there are decent arguments both for and against the existence of the item, and here, Lewis thinks, temperaments might matter.

Let us approach the problem with reference to Ramsey sentences. Suppose we are wondering whether ϕ exists. We construct a list of platitudinous desiderata of what we think ϕ is like: "ϕ is P_1," "ϕ is P_2," and so on. (Of course, the sentences do not need to be simple atomic predications, but it makes expression easier.) Let's say we come up with twenty such desiderata. We then conjoin these sentences, replace all mention of ϕ with the same variable, and bind that variable with an existential quantifier:

$$\exists x\,(x \text{ is } P_1 \,\&\, x \text{ is } P_2 \,\&\, \ldots x \text{ is } P_{20})$$

This allows us to say that ϕ exists if and only if the existential claim (the Ramsey sentence) is true. There are two ways in which it can be false: If there is nothing that satisfies all twenty predicates, or if there is more than one thing that satisfies all twenty predicates. (See Lewis 1970.)

But one should not be too strict about this, for there is a requirement to accommodate the fact that we can be mistaken about certain qualities of ϕ without it automatically following that ϕ does not exist. We gave the name "Earth" to the thing we live upon and at one time reckoned it flat (or at least a good many people reckoned it flat); but the discovery that the thing we live upon is a big ball was not taken to be the discovery that we do not live upon Earth. It was once widely thought that gorillas are aggressive brutes, but the discovery that they're in fact gentle social creatures was not taken to be the discovery that gorillas do not exist. Examples abound. We make mistakes, big and small, without facing skepticism at every turn.

Applying this to Lewis's argument: One of the desiderata for "value" pertains to *non-contingency*. (I shan't pause for further specificity.) But nothing that satisfies the

other desiderata (some of which can be systematized into *that which we desire to desire*) also satisfies non-contingency, and this is why Lewis claims that "genuine values would have to meet an impossible condition." But, he thinks, there are imperfect claimants—ones that satisfy *nearly* all that we want. Just as a spherical earth was an imperfect claimant of all that we might have wanted to say about the Earth (but close enough to avoid Earth skepticism), and gentle gorillas were imperfect claimants of all that we might have wanted to say about gorillas (but close enough to avoid gorilla skepticism), so too, Lewis thinks, his dispositional theory of value is an imperfect claimant of all that we might have wanted to say about value, but close enough to avoid moral skepticism.

But Lewis realizes that his dispositional theory is not close enough to silence all debate on the matter; it remains in a gray area. And the grayness persists because we have no methodology for making decisions concerning indeterminate cases. Lewis writes: "What it takes to deserve this name, not perfectly but well enough, was never settled" (1989: 136); and: "When it comes to deserving a name, there's better and worse but who's to say how good is good enough?" (137). Who indeed?

Elsewhere I have toyed with the idea that what determines this matter is a complex counterfactual about how we would, if called upon, as a matter of fact decide (Joyce 2006a: 201).[3] But this idea was never meant to help us out as a decision procedure, and it is, in any case, problematic. Who, exactly, is this "we," and in what circumstances are we being called upon to make this decision? Since I harbor no optimism that in making such decisions we collectively or individually follow any particular rules, overt or tacit, I'm inclined to think that our pronouncements on such matters can be influenced by the most arbitrary of stimuli (such as the plot of a popular movie). Thus in circumstances A we would decide that we've had a false belief about ϕ, correct that error, and carry on believing in ϕ (that is, we would be ϕ naturalists); while in circumstances B (which are not dramatically different from A) we would decide that our belief in the existence of ϕ has been a huge mistake (that is, we would be ϕ skeptics). This to some extent problematizes the counterfactual "If we were to decide on the matter, we would decide thus-and-so."

Subsequently, I suggested that what determines the question is how we *use* the concepts (Joyce 2006a: 201; 2007b: 65). Roughly: If concept ϕ—understood as satisfying twenty desiderata—has been used in ways U_1, U_2,… U_n, but the best imperfect claimant (call it ϕ^*, which satisfies, say, eighteen of those desiderata) *cannot* be put to the same range of uses, then this is sufficient grounds for denying that ϕ^* is "close enough" to count as a revision of ϕ.

I continue to think that this appeal to usage promises to make some headway toward solving some of the Lewisian *Is-it-close-enough?*-type questions. Regardless of whether one's temperament favors Lewis's "calm and conservative" preference for dispositional

[3] I have on occasion made what I take to be the equivalent point via a thought experiment concerning translation. (See Joyce 2001: 3, 26–7; 2006a: 71.)

moral naturalism, if it turns out that the dispositional property in question just isn't going to pull its pragmatic weight in everyday life—if it can't do the work of a full-blooded all-you-could-want *moral* property—then this would count decisively against this form of moral naturalism. Likewise, irrespective of whether one's mischievous temperament favors error-theoretic moral skepticism, if it turns out that Lewis's dispositional property *can* be pressed into service for all the practical roles that moral concepts play, then the day would go to the moral naturalist and his imperfect claimant.

One interesting and possibly surprising consequence of conceptualizing the problem in this manner is that it makes the debate between the moral naturalist and the moral error theorist at bottom an *empirical* debate. We ask, first, "What is morality used for?" Then we ask "Could concept such-and-such [the imperfect claimant] be used in this manner?" Both are difficult and complex matters, but they are both empirical questions, and the former, at least, seems reasonably tractable. Yet the remaining problems are more than just the difficulty of our gaining epistemic access to these empirical truths. The problem is that the indeterminacy simply reiterates in a new place. Suppose we have used concept ϕ for ten purposes—U_1, U_2,… U_{10} (idealizing horribly here, of course)—and suppose that the best imperfect claimant (call it ϕ^*) can be used in, say, eight of those ways. We cannot use ϕ^* for *everything* that we used to use ϕ for, but we can use it for *most* things. Well, is that close enough? I feel that at this point we can only reiterate Lewis's question: "Who's to say?" If we can find no good answer, then we have made not only the surprising discovery that the debate between the moral naturalist and the moral skeptic bottoms out in empirical disputes, but a second surprising discovery that there may be no fact of the matter about who is correct.

Carnap's Pragmatism

Lewis is evidently content to live with this indeterminacy between moral naturalism and moral skepticism. But we shouldn't give in just yet; there is an obvious avenue to survey: an appeal to pragmatism. If there is an indeterminacy in what we mean by "value" (say)—in that both the moral skeptic and the moral naturalist make defensible claims about what is necessary or sufficient for something to be a value—then let the matter be decided by which is most practically expedient; let us choose a precisification on the basis of which best serves our purposes.

Many philosophers, it seems to me, are drawn to this method of selecting sides when indeterminacy threatens, and, moreover, the common assumption regarding the present case is that an appeal to pragmatism will favor the moral naturalist over the skeptic. My aim is to call into question both moves. First, the appeal to pragmatism is fraught with difficulties and does not obviously represent an escape route from indeterminacy. Second, I think the assumption that an appeal to pragmatism (were it to prove workable for this purpose) would favor the moral naturalist can be subjected to pressure. But before arguing these claims I want to bring the noncognitivist on the stage, and I will do so via a discussion of Rudolf Carnap. Carnap's views are useful for our purposes,

since as well as advocating noncognitivism he espouses an interesting pragmatic methodology.

That the Carnapian notion of *explication* is relevant to our discussion is immediately evident from the characterization Carnap gives it in *Logical Foundation of Probability*: "By the procedure of explication we mean the transformation of an inexact, prescientific concept, the explicandum, into a new exact concept, the explicatum" (Carnap 1950a: 3). The test of a good explication is not whether it is "true" or "correct" (Carnap says such claims make "no good sense" (4)); but rather "whether the proposed solution is satisfactory" (1950a: 3)—where this satisfactoriness is in part a pragmatic notion, centrally depending on whether the proposed explicatum is fruitful and simple.

The pragmatism embodied in his notion of explication is plain elsewhere in Carnap's philosophy. In "Empiricism, semantics, and ontology" (1950b) he analyzes many traditional philosophical questions as "external questions." The question "Does the world exist," for example, is really a practical question of whether we should adopt the "thing language" according to which objects exist at a space and time within the world. In determining the practical payoff, "[t]he efficiency, fruitfulness, and simplicity of the use of the thing language may be among the decisive factors" (208). He says the same about the external questions "Do propositions exist?," "Do numbers exist?," and so on. According to Carnap, these questions should be interpreted as: "Are our experiences such that the use of the linguistic forms in question will be expedient and fruitful?" (213).

Carnap thus provides us with two ways that we might try to break out of the indeterminacy surrounding "value" (and other moral terms). First, we might try to *explicate* the notion, imposing precision on an indeterminate concept in a manner guided by expediency. Second, we might treat the question "Do values exist?" as an external question, deciding whether to employ the "value language" on the basis of its utility as an instrument.

Before examining these possibilities, we should have Carnap's own metaethical view before us. His years as a leading member of the Vienna Circle had made a moral skeptic of him—but of the noncognitivist rather than the error-theoretic variety. By 1935 he was staunchly advocating a prescriptivist version of noncognitivism: "Most philosophers have been deceived [by grammatical form] into thinking that a value statement is really an assertive proposition and must be therefore true or false . . . But actually a value statement is nothing else than a command in a misleading grammatical form . . . It does not assert anything" (Carnap 1935: 24–5). His noncognitivism had first become evident in 1929, in a lecture delivered to the Bauhaus in Dessau, where (according to fragmentary notes) he claimed: "Valuation (*Wertung*) is not the cognition of a fact but a personal attitude" (cited in Mormann 2007: 133). Interestingly, in his *Aufbau*, a younger Carnap had espoused a very different position, where values are a part of the world order, whose presence may be known and asserted (see section 152 of Carnap [1928] 1967). In what follows, however, I shall take the mature Carnap's noncognitivism

as representative of his view. (See Mormann 2007, for discussion of Carnap's change of mind.)

When Lewis outlines the range of options (in the long passage quoted earlier), the noncognitivist is noticeably absent; all that is on his radar is a continuum of positions between the moral error theory and moral naturalism. It is clear why this is so. Lewis's examination of "value" as a verb (roughly: *desiring to desire*) has provided him with license to treat "value" as a noun (roughly: *that which we are disposed to desire to desire*), which in effect is license to exclude noncognitivism. Thus the question that exercises him at the close of the paper is whether the world provides a property that is close enough to deserve this noun (albeit imperfectly). The noncognitivist, by contrast, will not let Lewis get that far. The noncognitivist endorses an account of "value" as a verb that does *not* provide license for treating "value" as a noun. Focusing on "valuing" as a linguistic act rather than a kind of mental attitude, Carnap tells us that value utterances are really commands, in which case the question of to what kind of entity these value utterances refer simply does not arise. When one says "X is good," Carnap thinks, one is saying something along the lines of "Pursue X!" The *goodness* has evaporated on this analysis. Asking whether the world provides anything close enough to deserve the noun "goodness" is, in Carnap's opinion, a pseudo-question, because concepts like *goodness* are pseudo-concepts and therefore not fit for analysis.

If there is no concept of *value*, then *a fortiori* there is no such concept to be made precise via explication. That is not to say, however, that explication can play no role in Carnap's metaethics. The concept of *value* accused of being an inexplicable pseudo-concept is the one associated with the noun "value." But Carnap certainly believes in the human activity of *valuing*, and the related concept (the one associated with the verb "to value") will not also be accused of "pseudo-ness." Moreover, Carnap also believes that human valuing will often take the form of issuing statements, and thus there is a noun ("value statement") whose associated concept is available for analysis and, if necessary, explication. On the only occasion that Carnap does explicitly sketch an explication of something normative, the explicandum is, indeed, "value statement" (Schilpp 1963: 1009ff).

Now, it is not my intention to discuss Carnap's proffered explication in detail, but rather make some general comments about it. First, it is interesting to note that he does not take himself to be deriving the distinction between noncognitive/evaluative utterances and descriptive/factual statements from any examination of natural language practices; rather, he is speaking of "possible kinds of meanings and the relations between these meanings" (Schilpp 1963: 1003). His aim is to establish that a noncognitive language is *possible*, and "to use it as a basis for the philosophical discussion of value problems" (1003).

This last comment reveals that Carnap's eye is, as usual, on the pragmatic payoff, but one should be aware of his restricted view on this matter. The "fruitfulness" of explication is understood in terms of the explicatum being "brought into connection with other concepts on the basis of observed facts; in other words, the more it can be used

for the formulation of laws" (Carnap 1950a: 6). Thus, it is not any general psychological kind of practical usefulness that is being claimed for noncognitivism; indeed, Carnap opines that whether one is a noncognitivist or a moral naturalist "is relatively unimportant in its influence of practical life" (in Schilpp 1963: 82).[4] When he does claim some advantage for his noncognitive explication of a value statement over any cognitive rival, it is in terms of a very slight (almost, one might say, unnoticeable) additional discriminatory power in the former language: The noncognitive language can provide an interpretation of a certain sentence of the cognitive language, but not vice versa (Schilpp 1963: 1004). However, the argument is unpersuasive, to say the least. Among its weaknesses is the fact that it claims this advantage regarding just one type of comparison; but who knows what advantages the cognitivist language might enjoy over the noncognitive language if we looked further afield?[5]

Let us conclude, then, that Carnap leaves the matter undecided. And he would, I think, agree with and even welcome this diagnosis. His Principle of Tolerance allows metalanguages to compete freely for our allegiance, to be judged not according to "truth" but according to fruitfulness.[6] Permit the noncognitivist and the cognitivist to advocate their respective languages; Carnap is tolerant of the competition. The language that best suits our theoretical purposes will be the eventual victor. "Let us grant to those who work in any special field of investigation the freedom to use any form of expression which seems useful to them; the work in the field will sooner or later lead to the elimination of those forms which have no useful function" (1950b: 221).

I will sum up before moving on. I set out to explore the possibility of an irresolvable indeterminacy between moral naturalism and moral skepticism. Lewis located one potential node of indeterminacy: between moral naturalism (of a certain kind) and error-theoretic moral skepticism. This is an indeterminacy over "value" as a noun. But before we get that far, we face the possibility of another node of indeterminacy: between cognitivism (which includes all forms of moral naturalism) and the other kind of moral skepticism: noncognitivism. This can be thought of as indeterminacy over "value" as a verb. (Quite possibly there is indeterminacy in the rival noncognitivist construals of *valuing*, too—between, say, Hans Reichenbach's volitionism and A. J. Ayer's emotivism.) Later I will further discuss the possible indeterminacy between naturalism and skepticism, but first I will investigate the potential undecidability of the debate between cognitivism and noncognitivism.

[4] Carnap does think that clearly demarcating value questions from factual questions is practically important, and that the failure to do so "leads to confusions and misunderstandings in discussion of moral problems in personal life or in political decisions" (in Schilpp 1963: 81)—but that is not the same as thinking that the distinction between noncognitivism and cognitivism makes a practical difference.

[5] A couple of other weaknesses with Carnap's arguments are the following. First, he seems to be concerned only with showing how purely noncognitive sentences ("pure optatives," as he calls them) are possible; but the moral cognitivist need not deny this, but rather simply denies that *moral utterances* are instances of purely noncognitive sentences. Second, Carnap has only a very restricted kind of cognitivism in his sights when he performs this comparison.

[6] They are *metalanguages* because the Principle of Tolerance is applied after sentences in the material mode of speech have been translated into sentences in the formal mode.

Cognitivism versus Noncognitivism

Carnap, as we saw, invites open competition with his cognitivist rival. But the mere fact that one is sufficiently undecided between two theories of X (or sufficiently sportsmanlike) to be willing to allow them to "compete" in some manner for victory hardly means that there is a substantive indeterminacy about X. For there to be indeterminacy it must be the case that (a) this competition yields no winner, and (b) no other form of decision procedure is available. Could there really be no fact of the matter about metaethical cognitivism versus noncognitivism?

One might think that of course there is a fact of the matter. This debate is, essentially, about what we are doing when we *value* something, and the act of valuing is a phenomenon available to our investigation. This is why although there are plenty of advocates of an error theory about *values*, there are no error theorists about *valuing*.[7] We do not doubt that valuing occurs, and this type of action has certain features which determine the answers to various questions we might ask about the phenomenon. So surely the debate between the cognitivist and the noncognitivist must have an answer, even if we are not sure yet what it is.

But this line of reasoning is mistaken. Valuing may well be an actual phenomenon displayed before us, available to our scrutiny, but the phenomenon will yield specific answers only if we ask specific questions. There are different ways of framing the cognitivism/noncognitivism question, but the standard way is: "Are moral judgments assertions?" Here *valuing* is treated as a linguistic activity, and the question concerns what kind of speech act constitutes the activity. Specific enough?

Not necessarily. Questions arise as to what is meant by "assertion." There are significant differences among the views of assertion advocated by C. S. Peirce (1934), Michael Dummett (1959), J. L. Austin (1962), John Searle (1969), Michael Slote (1979), Robert Brandom (1994), and Timothy Williamson (2000)—to name just a few. Can these differences be settled? Perhaps not. Perhaps different accounts of *assertion* work better in different theoretical environments, with there being no decisive facts to appeal to in order to settle either which is the best account of *assertion* overall or which is the best account of *assertion* to apply to the moral case.

Many accounts of assertion state or imply that the person making an assertion aims at the truth, or aims to state a fact (see Dummett 1981: 300; Williams 1966). So when someone claims "Stealing is wrong," is he or she stating a fact? It depends what you mean by "fact"—of which, it comes as no surprise, there are competing theories. One might distinguish between "fact" in a metaphysically robust sense, and "fact" in some more minimalist sense. This, in turn, would give us a distinction between "assertion" in a robust sense and in a minimalist sense. Indeed, we find exactly this kind of distinction

[7] Unless, of course, it follows from the endorsement of some grander skepticism, encompassing doubt about cats and dogs, furniture, other people, and so on. Even an error theorist about all mental phenomena need not be an error theorist about valuing, since one can treat valuing as a species of linguistic act or behavioral response.

appearing in metaethical debates over quasi-realism. The quasi-realist eschews the metaphysics of moral realism but thinks that language with all the trappings of realism is still permissible (Blackburn 1984, 1993). Thus, even if there is no metaphysical property of *wrongness*, the quasi-realist is happy to carry on claiming not only "Stealing is wrong," but "'Stealing is wrong' is true," "There is a property of *wrongness* (and stealing has it)," "Stealing is wrong regardless of my attitude," "I hereby assert that stealing is wrong," and "I really, really mean it." A typical quasi-realist (noncognitivist) view is that while we can claim all these things (truth, fact, property, assertion) for morality in a *minimalist* sense, we cannot do so in a *robust* sense. In this way the quasi-realist hopes to vindicate the realist trappings of moral discourse while continuing to distinguish the position from moral realism.

This situation is not simply one of philosophers arguing over the correct analysis of a concept whose extension all will substantially agree to; rather, the extension may very well differ significantly depending on which conception is preferred. It is possible that the question "Are moral judgments assertions?" should receive a positive answer if using a minimalist conception of *assertion* but a negative answer if using a robust conception. But nor is this situation simply one of an ambiguous concept, for the parties to the dispute will usually insist that their preferred conception is uniquely correct. We will not, for example, typically find moral realists accepting that the minimalistic notion of *assertion* is legitimate for certain contexts; rather, they will generally despise the very idea (along with the minimalistic versions of truth, fact, and so on). Similarly, the quasi-realist may well claim that the robust version of *assertion* is a mere philosopher's fancy, and perhaps not a well-formed idea at all. Perhaps this dispute could one day be settled; perhaps one of the disputants is, after all, uniquely correct. But it is also possible (though I am not arguing that it is the case) that neither is uniquely correct; perhaps there is nothing that determines the correctness of one conception over the other; perhaps the concept of *assertion* is simply indeterminate across the relevant cognitivist/noncognitivist gap.

I want now to return to the possibility of settling this dispute by an appeal to convenience and expedience, though this time I propose to understand these matters (albeit vaguely) in a broader, more psychological, sense than Carnap intended. My worries with this suggested decision procedure arise before we even begin to wonder about what the practical upshot of moral noncognitivism is. My concern arises because the question "Expedient to whom?" seems a perfectly fair yet unanswered response.

Many pragmatist suggestions speak in a cavalier fashion about "what is useful to our practical purposes" without pausing to wonder to whom the "our" refers; the common background assumption is that there is a convergence in practical requirements. But this assumption is, upon reflection, wildly implausible. What if noncognitivism suits some people while cognitivism suits others? Then one precisification of "assertion" (say) will be practically expedient for some, while an opposed precisification will be practically expedient for others. This is reminiscent of an old complaint against the Pragmatist school of philosophy, one that Bertrand Russell heatedly made in his essay

on the topic (and on every possible occasion thereafter): "One gathers . . . that a Frenchman ought to believe in Catholicism, an American in the Monroe Doctrine, and an Arab in the Mahdi" (Russell 1910: 97). That Russell had an accurate or fair image of his target is unlikely (see Haack 1976), but here my worry concerns nothing so grand as Pragmatism as a school of philosophy or a theory of truth. My target is just one kind of appeal to pragmatic considerations in the hope of settling a philosophical dispute, and here, it seems to me, Russell's question remains a live one.

Regarding the practical upshot of noncognitivism's being true or false, is it really reasonable to suspect that there will be divergence among individuals' interests? Alasdair MacIntyre's analysis of the Bloomsbury Group might serve as an illustration (MacIntyre 1981). MacIntyre interprets the Bloomsbury circle as developing a peculiar moral discourse: Surrounded by a community that employed moral vocabulary in an assertoric way, the members of the Bloomsbury Group (unwittingly) did otherwise. Impressed with G. E. Moore's elevation of beauty, love, and knowledge to the pedestal of "intrinsic moral values," this group of intellectuals used moral language as a manipulative tool to persuade others of the delicacy (and thus superiority) of one's own aesthetic sensibility. John Maynard Keynes recalls that about these matters it was "useless and impossible to argue" (Keynes 1949: 85). He goes on:

> In practice, victory was with those who could speak with the greatest appearance of clear, undoubting conviction and could use the best accents of infallibility. Moore at this time was a master of this method—greeting one's remarks with a gasp of incredulity—*Do* you *really* think *that*, an expression of face as if to hear such a thing reduced him to state of wonder verging on imbecility, with his mouth wide open and wagging his head in the negative so violently that his hair shook. *Oh!* He would say, goggling at you as if either you or he must be mad; and no reply was possible. (85)

MacIntyre's thesis that noncognitivism was true of the Bloomsbury Group while not true of the wider populace is an interesting but knotty claim—but it is not, in any case, our present concern. Rather, I am suggesting that it might have *suited the Bloomsbury Group's purposes if noncognitivism were true*—because, roughly, it would *validate* their practices—while not serving the purposes of others. Thus the proposal that we should settle any indeterminacy between noncognitivism and cognitivism—and thus between moral skepticism and moral naturalism—by reference to "what is in our practical interests" is seen to be seriously problematic.

The Benefits of being Horribly Wrong

The pragmatic payoffs of noncognitivism versus cognitivism may be inscrutable and divergent, but surely we can hope to do better at the other node of potential metaethical indeterminacy: moral naturalism versus the error theory? Here the naturalist might feel more confident that there is nothing to be said, pragmatically, in favor of the error-theoretic position. Here the naturalist might feel satisfied with Lewis's diagnosis,

assured that if all that lies between moral naturalism and moral skepticism is "temperament," then he, the naturalist, may declare victory.

However, in my opinion, if the argument has come to this point, the moral naturalist has no particular grounds for complacency, for there are, I think, considerable *pragmatic* claims to be made in favor of the moral error theory.

The moral error theory tells us that we have been massively mistaken about the world, about ourselves, and about the relation between the two. Finding out that one has been massively mistaken is, it can be assumed, an uncomfortable state, and being uncomfortable is unpleasant. Finding out, moreover, that our mistake has been as great as the moral error theorist would have us believe—ubiquitous across all human societies, through all history, pervading nearly every aspect of our social existence—is likely to make us uncomfortable to the point of positive distress. But one needs to look deeper. Being uncomfortable can be a useful state (after all, pain serves a good biological purpose), and distress can, ultimately, be instructive.

By way of leading you into my thinking, consider what we have learned from experimental psychology about confabulation. There are various kinds of cognitive impairment (to memory, to perception, and so on) for which the subject will compensate by creating a false narrative to "fill in the gaps." Someone suffering from Korsakoff's syndrome fails to admit, even to herself, the severe memory loss she suffers, but rather invents elaborate stories to cover her confusion. Asked why she is in the hospital, for example, the patient might genuinely assert that she is visiting someone or applying for a job there. Sometimes stroke patients with partial paralysis will deny the paralysis, inventing sincere excuses for why that part of their body cannot perform actions. (See Hirstein 2005; Schnider 2008.) Fascinating as these kinds of case study are, they are really just the lurid side of the phenomenon of confabulation, much of which is far more routine and commonplace. We all superimpose justifications and explanations onto our actions when the real sources of motivation are hidden from us; we all string together coherent life narratives from fragmented and distorted memories; we all confabulate. (See Wheatley 2009.)

This is a disconcerting truth when its full extent is apprehended ("I am confabulating nearly all the time!"). It forces us to adopt a new perspective on who we are and how we work. Experimental psychology can bring us to see that our intuitive everyday image of ourselves and our lives is, in many ways, badly mistaken.[8] Yet, I claim, this unsettling apprehension is often a good thing. It is good not merely because it is *true*, but because it is useful. Of course, we tend to compartmentalize this analytical awareness, and quickly return to confabulating as soon as we relax our attention. And that's fine. What I am speaking in favor of is the value of regularly returning to the critical

[8] Of course, it is not just the phenomenon of confabulation that shows this. The broader claim that experimental psychology has revealed that we are, in a multitude of ways, routinely mistaken about what is really going on with our motivations, emotions, and belief-formation processes, is far less contentious and far too well documented to require citations.

perspective from which we realize that we are more opaque and baffling beings than everyday life presupposes.

Consider this scenario. You are riding in a bus through a city, absent-mindedly watching pedestrians and street scenes, when the thought strikes you: "What odd creatures we are." And suddenly you undergo a Desmond-Morris-style gestalt shift: You think of the pedestrians under the description "hairless upright social apes"—you literally see them *as* hairless upright social apes. Suddenly the city appears like a swarming nest of primates, commonplace urban activities come to seem like bizarre arbitrary rituals, the projects and concerns that ordinarily occupy our minds are seen as weird and futile. And perhaps the oddest thought of all is: "Hey, *I'm* one of those naked apes, too!"

When one thinks along these lines, it's not that one all of a sudden forms new beliefs, for we all know (those of us that are sensible, at least) that we are hairless upright social primates, with bodies and minds shaped by Pleistocene Africa.[9] But this knowledge is compartmentalized during our everyday lives; it takes an effort to attain the critical distance to *feel* the truth of such beliefs, to be dumbstruck at the sheer strangeness of it all.

And that feeling—a feeling of alienation from the customary images of oneself and the society one inhabits—is, I contend, a valuable perspective to adopt. Like a full-blooded encounter with the aesthetic sublime, it is not an entirely pleasant experience. The sublime can make one weak in the knees with confusion and awe. Yet we seek out such encounters (sometimes) because in that feeling we sense that we are in touch with something authentic, as if we have stepped back from our familiar humdrum concerns and tasted the enormity of the universe and the fragility of our place in it.

This is the benefit I would like to advertise of seeing ordinary beliefs as horribly wrong. It is good to be epistemologically shaken. It reminds us of how shockingly ignorant we are and how mysterious everything really is. It is both a corrective to epistemic complacency and a spur to intense reflection and inquiry. Discovering that everyone is horribly wrong about something fundamental to our lives is not just sobering and valuably humbling, but intellectually exciting. Endorsing an error theory puts hairs on your chest, epistemologically speaking.

Well, that's just me. If the last few paragraphs fail to speak to you, then that's okay; that, indeed, is my central point. I am not claiming that every person would benefit from recognizing his/her profound error and therefore we should all be error theorists. Remember that the present dialectic presupposes that there are already some decent arguments in favor of the moral error theory (and decent arguments for moral naturalism), so the pragmatic considerations just raised are supposed to supplement solid arguments, not stand in place of them. But nor am I advancing the amended thesis that

[9] This claim that we have "minds shaped by Pleistocene Africa" should not be over-interpreted as an endorsement of any strong form of Evolutionary Psychology. The prehistorically forged human mind may be designed to be a plastic and generalized problem solver.

every person would benefit from recognizing his/her profound error and therefore *if we have decent arguments in favor of the error theory* then we should all be error theorists. I am in fact nowhere claiming that every person would benefit from recognizing his/her profound error. Perhaps some people would and some people would not. My point is (again) that even in the case of moral naturalism versus the moral error theory, the pragmatic cost-benefit analysis is complex, and *there is no discernible reason to expect convergence*.

I can imagine critics complaining that even if there do exist such intellectual benefits to be gained from an apprehension of the depths of human epistemic fallibility, surely they pale beside the losses that will be incurred from an endorsement of a moral error theory. Surely (the complaint goes) a belief in the moral error theory will lead to a breakdown of good citizenship, to a loss of motivation to enter sincerely into cooperative ventures, to cheating one's friends when the chips are down, to selfishness and debauchery, and so forth. But these vague fears are expressed far more often than any scrap of evidence in their favor is offered. While I agree that moral belief contributes to an individual's "cooperative motivation" (to choose an umbrella term) (see Joyce 2006a: chapter 4), the idea that moral belief is the *only* thing that does so is a hypothesis as doubtful as it is depressing. And even if the hypothesis were true, the claim that a breakdown of cooperative motivation is the inevitable outcome of embracing the moral error theory implies the similarly doubtful hypothesis that moral belief is the only thing that *can* motivate cooperation. No, there is no evidence that moral skepticism leads to crimes and misdemeanors. I have met many moral error theorists, and am yet to encounter one who wasn't perfectly civilized or was any less trustworthy than anyone else. By contrast, if one reflects upon the worst mass criminals that humanity has produced, one is likely to encounter not skeptics but individuals moved by moralistic fervor. (See Garner 2010; Hinckfuss 1987.)

Of course, these observations don't settle the matter. It is entirely possible that, despite my anecdotal evidence, in fact moral error theorists are always less flourishing and always inferior citizens when compared to moral believers. Or perhaps it goes the other way round. What I am seeking to establish is that we shouldn't jump to any conclusions regarding these empirical matters, and in particular shouldn't assume that there will be convergence among individuals one way or the other. Though there may well be practical costs to being a moral error theorist, I am content on this occasion to attempt to publicize some of the neglected practical benefits.

The idea that our interests are best served by a methodology that always favors theories which mesh with our common sense is, for my money, shortsighted and disappointing. Perhaps some individuals require that kind of constant reassurance—perhaps many do—but not all do. Some find the prospect that common sense is massively mistaken an object of contemplation that is, while anxiety-provoking (partly *because* anxiety-provoking), profoundly rewarding. Given the choice between a theory that vindicates some aspect of common sense and a theory that overturns it, some people will find the latter more valuable. I cannot resist here quoting Bertolt Brecht's

gloriously scathing assessment: "I'm not writing for the scum who want to have the cockles of their hearts warmed" (Willett 1957: 14).

Quine's Sectarianism and Ecumenicalism

I have argued that the debate between the moral naturalist and the moral skeptic—which I broke down into (i) that between the moral cognitivist and the moral noncognitivist, and (ii) that between the moral naturalist and the moral error theorist—may be undecidable. Key concepts like *assertion* or *value* may be sufficiently indeterminate to permit all aforementioned theoretical possibilities. The prospect of avoiding this conclusion by an appeal to pragmatism has been explored and rejected; pragmatic solutions presuppose an implausible convergence in practical interests. It should be stressed that I have not argued that we *do* face this indeterminacy, only that we might. For all I have said here, it is possible that the moral naturalist will triumph *tout court*. Or perhaps the moral skeptic will.[10] But let us suppose for now that these or similar arguments favoring undecidability are sound. Where do we go from here?

Having discussed Lewis and Carnap, it seems apt to end with the philosopher who best connects them: W. V. Quine.[11] Quine famously argues for the underdetermination of scientific theory: that logically incompatible global theories may be equally consistent with all possible evidence and equally satisfactory with respect to internal theoretical virtues (such as simplicity and clarity) (Quine 1960, 1975). As to the "Where do we go from here?" question, Quine vacillated. In certain moods he espoused a *sectarian* response, according to which we should continue to endorse our favorite theory as robustly as ever, rejecting all alternatives as false or meaningless (Quine 1981: 21, 1986). But on other occasions he preferred an *ecumenical* response, whereby we accept that indeterminacy reveals all satisfactory though incompatible theories to be true (Quine 1981: 29, 1989).

We face a similar decision regarding metaethical indeterminacy. Suppose, like Lewis, one is inclined to endorse a form of moral naturalism. And suppose the possibility that I have discussed turns out to be a reality: that another person could, with equal legitimacy, endorse some form of moral skepticism. What attitude should the first person take—to push on with moral naturalism, declaring moral skepticism bankrupt (sectarianism), or perhaps to take a more tolerant view of alternatives (ecumenicalism)? The moral skeptic, of course, faces the same dilemma.

[10] Readers might be aware that I have argued for the error theory in the past (Joyce 2001, 2006a, 2011a, 2011b), and I dare say that in the future I will return to doing so. Despite my sympathetic attitude toward Lewis's pluralistic position, I have not officially conceded that his "imperfect claimant" (or any other contender) is close enough to warrant our endorsement as a *morality*. But even if I were to make that concession, continuing to argue for an error theory at the level of metaethics is consistent in every way with favoring pluralism at the metametaethical level. (See note 12 following.)

[11] Carnap taught Quine; Quine taught Lewis.

Quine became aware of this conflict in his work, and tended to plump for sectarianism. He even inserted comments into later editions of *Theories and Things* in order to distance himself from the ecumenical noises from his earlier self. He diagnoses his own indecisiveness on this matter by saying: "The fantasy of irresolubly rival systems of the world is a thought experiment out beyond where linguistic usage has been crystallized by use" (Quine 1990: 100).

Faced with the nodes of indeterminacy highlighted in this essay, I would counsel neither sectarianism nor ecumenicalism in particular, but rather what might be called "metaethical ambivalence."[12] This perspective begins with a kind of metametaethical enlightenment. The moral naturalist espouses moral naturalism, but this espousal reflects a mature decision, by which I mean that the moral naturalist doesn't claim to have latched on to an incontrovertible realm of moral facts of which the skeptic is foolishly ignorant, but rather acknowledges that this moral naturalism has been achieved only via a non-mandatory piece of conceptual precisification. (This describes Lewis's tolerant view.) Likewise, the moral skeptic champions moral skepticism, but this too is a sophisticated verdict: not the simple declaration that there are no moral values and that the naturalist is gullibly uncritical, but rather a decision that recognizes that this skepticism has been earned only by making certain non-obligatory but permissible conceptual clarifications.

This enlightened awareness of the legitimacy of one's opponent does not, however, suffice for metaethical ambivalence. For the next step we can turn again to Quine, recalling that even when advocating sectarianism he does not encourage adherents of one scientific theory to blinker themselves uncritically to other possibilities. What is at stake for Quine is the permissibility of applying the truth predicate to alternative theories to one's own; but even in those moods when he claims that one can call only one's own favored theory "true," he still allows that one can "oscillate between . . . theories for the sake of added perspective from which to triangulate on problems" (Quine 1990: 100). And this is, perhaps, indicative of the most sophisticated pluralism of all. The enlightened moral naturalist doesn't merely (grudgingly) admit that the skeptic is warranted in his or her views, but is willing sometimes to adopt the skeptical position in order to gain the insights that come from recognizing that we live in a world without values. And the enlightened moral skeptic goes beyond (grudgingly) conceding that moral naturalism is reasonable, but will sometimes assume that perspective in order to gain whatever benefits come from enjoying epistemic access to a realm of moral facts. Such a metaethical ambivalence will require a temperament that favors neither the conservative nor debunking attitudes mentioned by Lewis, but rather an intellectual courage to cut oneself loose from the comforting familiarity of assuming that there is always a fact of the matter.

[12] Broadly parallel to David Wong's *moral ambivalence* (2006: 20ff). I take it that just as Wong's moral ambivalence is ambivalence *about* moral systems—an ambivalence that occurs *at* the metaethical level—so too my metaethical ambivalence is ambivalence *about* metaethical theories—an ambivalence that occurs *at* a metametaethical level. For insightful discussion of doxastic ambivalence, see Roorda (1997).

PART II

Evolution and Debunking

5

The Origins of Moral Judgment

Introduction

Is human morality a biological adaptation? And, if so, should this fact have any substantial impact on the ethical inquiry of how we should live our lives? In this essay I will address both these questions, though will not attempt definitively to answer either. Regarding the former, my goal is to clarify the question and identify some serious challenges that arise for any attempt to settle the matter one way or the other. Regarding the latter, my ambitions here are restricted to some brief critical comments on one recent attempt to answer the question in the affirmative.

Let us start with Darwin:

> I fully subscribe to the judgment of those writers who maintain that of all the differences between man and the lower animals, the moral sense or conscience is by far the most important… [A]ny animal whatever, endowed with well-marked social instincts, the parental and filial affections being here included, would inevitably acquire a moral sense or conscience, as soon as its intellectual powers had become as well, or nearly as well developed, as in man. (Darwin [1879] 2004: 120–1)

There are several features of this passage worth highlighting. First, the trait that is under discussion is described as "the moral sense or conscience," which, it seems safe to claim, is a faculty that produces moral judgments. Darwin is not here wondering whether being morally good is the product of evolution, but rather whether the capacity to make self-directed moral judgments is the product of evolution. A moment's reflection on the myriad of ways in which morally appalling behavior may be motivated by a sense of moral duty should suffice to illuminate the distinction.

The second conspicuous feature of the passage is that Darwin sees the moral sense as emerging (inevitably) from other traits: "social instincts" combined with "intellectual powers." The latter powers he goes on to mention are memory, language, and habit. This raises the possibility that Darwin does not see the moral sense as a discrete psychological adaptation but rather as a byproduct of other evolved traits. In fact, he appears wisely to steer clear of adjudicating on this matter. When focused on the social instincts generally (rather than the moral sense in particular), he writes that "it is… impossible to decide in many cases whether certain social instincts have been acquired through natural selection, or are the indirect result of other instincts and faculties" ([1879] 2004: 130).

Contemporary debate among philosophers (in particular) over whether the human moral sense is an adaptation has not always been so cautious. Several recent authors have developed arguments to the conclusion that human moral judgment is not a discrete adaptation but rather a byproduct of other psychological traits (Nichols 2005; Prinz 2008a; Ayala 2010; Machery and Mallon 2010). Let us call these people "spandrel theorists" about morality. Others, myself included, have advocated the view that the human moral sense is a biological adaptation (Alexander 1987; Irons 1996; Krebs 2005; Dwyer 2006; Joyce 2006a; Mikhail 2011). We'll call these people "moral nativists." My first substantive goal in this essay is to reveal how difficult it is to resolve this matter.

Adaptations versus Spandrels

The spandrel theorist proceeds by offering "non-moral ingredients"—themselves quite possibly adaptations—which are sufficient to explain the emergence of moral judgment. We have seen Darwin mention such things as language use, social instincts, and memory. Francisco Ayala emphasizes "(i) the ability to anticipate the consequences of one's own actions; (ii) the ability to make value judgments; and (iii) the ability to choose between alternative courses of action" (2010: 9015). Jesse Prinz (2008a, 2014) considers such non-moral ingredients as meta-emotions, perspective taking, and the capacity for abstraction. Here I will take as my exemplar the view of Shaun Nichols (2005), but the general point I shall make could be leveled at any of the aforementioned (and, indeed, against any spandrel theorist).

The two non-moral ingredients that Nichols focuses on are a capacity to use non-hypothetical imperatives[1] and an affective mechanism that responds to others' suffering. He writes that:

> . . . both of the mechanisms that I've suggested contribute to moral judgment might well be adaptations. However, it is distinctly less plausible that the capacity for core moral judgment itself is an adaptation. It's more likely that core moral judgment emerges as a kind of byproduct of (*inter alia*) the innate affective and innate rule comprehension mechanisms. (2005: 369)

An obvious way of critically assessing Nichols's claim would be to question whether these two mechanisms, working in tandem, really are sufficient to explain moral judgment (for the sake of simplicity I'm ignoring Nichols's sensible "*inter alia*" in the previous quote). This would involve describing the two mechanisms highlighted by Nichols in much more detail, searching for empirical evidence (for example, can an individual

[1] A hypothetical imperative ("Go to bed now," for example) recommends that the addressee pursue a certain means in order to achieve one of his/her ends (to get a good night's sleep). If it turns out that s/he lacks that end, then the imperative is withdrawn. A non-hypothetical imperative demands an action irrespective of the addressee's ends. For example, the imperative "Do not slaughter innocents" is not withdrawn upon discovery that the addressee loves slaughtering innocents, will not get caught, and does not give a fig for morality. Moral imperatives are a subset of non-hypothetical imperatives. Non-moral non-hypothetical imperatives include etiquette, road regulations, rules of games and sports, and the norms of institutions generally.

have one of these mechanisms impaired and yet still make moral judgments?), and so forth. But the question I want to ask is much more general: What determines whether a trait (i) is a byproduct of other mechanisms x, y, and z, or (ii) is an adaptation dependent upon pre-adaptational sub-mechanisms x, y, and z? Answering this question in the abstract is fairly straightforward, but having a procedure for empirically determining whether a trait is one or the other is considerably more difficult. Let me explain.

No psychological faculty for producing a species of judgment is going to exist as a monolithic entity that takes inputs and magically produces outputs; all such faculties will depend on the operation of numerous psychological sub-mechanisms, which in turn depend on sub-sub-mechanisms, and so on. Suppose that Nichols is correct that the two mechanisms he highlights are indeed sufficient to explain the phenomenon of moral judgment. One interpretation—the one Nichols favors—is that the capacity for moral judgment is a byproduct of the operation of these two mechanisms. But a second hypothesis is always available: that the capacity for moral judgment is a distinct adaptation of which these are two sub-mechanisms. The second hypothesis is true if (and only if) the manner in which these two mechanisms interact has been at all modified by natural selection because their interaction has some impact on reproductive fitness. Let us suppose first of all that these two mechanisms evolved for their own evolutionary purposes. But in certain circumstances they interacted, in such a way that the trait of moral judgment emerged as a byproduct. Suppose further, however, that this new trait (moral judgment) had some reproductive relevance, such that the process of natural selection began to "tinker"—perhaps strengthening the interaction of the two mechanisms in some circumstances, dampening it in others. If this has occurred, then the capacity for moral judgment is no longer a mere "byproduct" but rather an adaptation in its own right. (Of course, one can still maintain that it originally appeared as a byproduct, but this is true of virtually everything that counts as an adaptation; see Dennett 1995: 281.)

In sum, spandrel theorists about morality seem to think that it suffices to establish their view if they offer non-moral ingredients adequate to account for moral judgment. But the consideration just raised indicates that this matter is not so straightforward, for any spandrel hypothesis can be interpreted instead as a description of the sub-mechanisms of the nativist moral sense (and if the ingredients mentioned are indeed adequate to explain moral judgment, then so much the better for the resulting nativist hypothesis).

But how would one distinguish empirically between these two hypotheses? The difference between an adaptation and a byproduct cannot be discerned by consulting intrinsic features of the organism, no matter in what detail. Consider Stephen Jay Gould's architectural analogy that originally provided the term "spandrel" (Gould and Lewontin 1979). Renaissance architects faced the design challenge of mounting a dome upon a circle of arches; when this is accomplished, the spaces between the arches and dome produce roughly triangular areas of wall: spandrels. These areas of wall are not design features—they are byproducts of the design features. Yet one could not

discern this by examining the intrinsic structural features of the building; one must know something about the purposes of the architects. It is, after all, conceivable that an architect may have a direct interest in creating spandrels, in which case the dome and arches would be byproducts. The resulting church would be intrinsically indistinguishable from the ordinary church for which the spandrels are byproducts.

In the same way, in order to know whether a trait is an adaptation as opposed to a byproduct one must understand something of the intentions of the architect—in this case, the forces of natural selection that existed during the period of the trait's emergence. Lacking, as we usually do, concrete evidence of the subtle evolutionary pressures operating upon our ancestors, our epistemic access to this information will always depend to some extent on intelligent inference. Consider, for example, Nichols's contention that the capacity to use non-hypothetical imperatives is an adaptation whereas the capacity to use moral imperatives is a byproduct. An alternative view is that the capacity to use moral imperatives is the adaptation while the more general capacity to use non-hypothetical imperatives is the byproduct. One could not decide between these hypotheses simply by examining the human organism; rather, the decision would have to involve comparing the plausibility of two conjectural hypotheses. On the one hand, one might hypothesize that the ancestral environment contained adaptive problems for which the specific capacity to use moral judgments would be a reasonable solution. Alternatively, one might hypothesize that the ancestral environment contained adaptive problems for which the specific capacity to use non-hypothetical imperatives would be a reasonable solution. In either case, the adaptive problems would need to be described in a manner supported by available evidence. To the extent that the former hypothesis turned out to be more plausible than the latter, moral nativism would be supported. But if the latter were more plausible than the former, then support would be provided for the spandrel view. A troubling possibility, of course, is that we may very well find ourselves lacking solid ground for favoring either kind of hypothesis over the other, in which case we'd lack ground for claiming with confidence which trait is the adaptation and which the byproduct. One can see now, perhaps, the wisdom of Darwin's quietism on this matter.

What is the Trait under Investigation?

I have been outlining one way in which the dispute between the moral nativist and the spandrel theorist is likely to run aground. However, it might reasonably be responded that this problem is of little consequence, since the contrast that is of greater theoretical interest is whether the capacity to make moral judgments is the product of evolutionary forces (whether an adaptation or a byproduct) or is an acquired ability. Frans de Waal calls the latter position "veneer theory": the view that morality, along with cooperative and altruistic tendencies in general, is "a cultural overlay, a thin veneer hiding an otherwise selfish and brutish nature" (de Waal 2006: 6). I doubt that many people nowadays endorse the veneer theory; that humans have been designed by natural

selection to be gregarious and cooperative seems beyond reasonable doubt. The devil lies in the details of *how* we are gregarious and cooperative. Note that declaring that we are by nature gregarious and cooperative is not to declare in favor of moral nativism, for it remains entirely possible that our social nature consists of biologically entrenched tendencies toward altruism, sympathy, love, and so forth, while the capacity to make moral judgments is an acquired and relatively recent cultural characteristic.

This observation, however, focuses attention on the knotty question that lies at the heart of these debates: What is a moral judgment? There is little to be gained in arguing over whether a trait is an adaptation or a spandrel, innate or acquired, if we do not have a firm handle on the nature of the trait under investigation. It is a great inconvenience to these debates that the concept *moral judgment* is a slippery and highly contested idea even among those who are supposed to be experts on the topic: namely, metaethicists.

In order to approach this problem, let us pause to compare chimpanzee sociality with human sociality. De Waal has often claimed that chimpanzee life contains some of the "building blocks" of morality (1992, 2006). He focuses on such things as reciprocity, consolation behavior, inequity aversion, empathy, and the following of rules of conduct reinforced by others. At the same time, de Waal is positive that chimpanzees do not make moral judgments (1996: 209; see also Boehm 2012: 113–31). This raises the question of what additional building blocks need be added, or how the building blocks need be rearranged, in order to create something deserving of the name "a moral sense." The fact that the answer is not at all clear problematizes the whole dialectic concerning the evolution of morality. In what follows I will attempt to say something useful on the matter.

A striking feature of the chimpanzee building blocks is that they seem to require emotional arousal. A deviation from a social rule in chimpanzee society receives a negative response only because those giving the response get angry. Consolation behavior is provided only by those in whom sympathy has been stirred. A reciprocal act (grooming behavior, say) occurs because the reciprocator feels friendly and caring toward the recipient (or, perhaps, feels fearful of the reprisal that non-reciprocation might bring). What chimpanzees seem to lack is a psychological apparatus that could motivate such behaviors in the absence of emotional arousal. In humans, by contrast, a deviation from a social rule might receive a negative response because those giving the response judge that it is deserved; consolation behavior might be provided by those who consider it right to do so; reciprocation might be offered because one judges oneself to have a duty to repay a debt; and so forth.

There are many who would claim that because the prominent "building blocks of morality" seem to be affective phenomena, the fully-fledged moral faculty must also be an affective mechanism.[2] One might argue, for example, that what humans have

[2] Here I am using the terms "affective," "noncognitive," and "conative" synonymously. I am not shunning the term "emotional," but am treating it with care, for emotions—at least many of them—are mixtures of affective and cognitive components. (For this reason, I do not consider Christopher Boehm's claim that the

(which other primates lack) is the capacity to have meta-conations. Perhaps if an individual not only dislikes a certain behavior, but likes the fact that she dislikes it (and perhaps also dislikes anyone who fails to dislike it, and likes anyone who does dislike it) then we may speak of her "morally disapproving" of the behavior. Perhaps if one's dislike of another's action prompts not only anger, but a disposition to feel anger at those who do not also feel anger at the action, then we may speak of one's judging that the anger is merited (see Blackburn 1998: 9–13; Prinz 2007: 113–15).

I find this line of reasoning unpersuasive. The building blocks of morality found in chimpanzees (and, by presumption, our ancestors) may well be affective phenomena, but it is entirely possible that the crucial modification of these building blocks in the human lineage was the addition of certain cognitive aptitudes. After all, generally speaking, the explosion of cognitive abilities is surely the most striking aspect of recent human evolution. Moreover, it is far from obvious, just on conceptual grounds, that one can really build a moral judgment from these affective ingredients alone. The natural way of assessing the claim is to examine potential counterexamples, of which there are two types. First, can we imagine these noncognitive capacities being deployed without a moral judgment occurring? Second, can we imagine a moral judgment occurring without these noncognitive capacities being deployed? I'm inclined to think that the answer to both questions is "Yes."

Suppose I am strolling among a group of normally docile animals when one bites me aggressively. Naturally, I dislike this; perhaps I smack the animal on the nose in order to make it release me. Perhaps, moreover, I judge that it is important that these animals do not form aggressive habits (maybe my children often play in their vicinity), so I would wish to see others smack the animal if bitten. Perhaps I go so far as to dislike anyone who would not smack the animal if bitten. Yet these emotions and meta-emotions do not appear to amount to a moral judgment of the animal's behavior. It does not seem that I judge that the animal deserves to be smacked; indeed, I do not treat the animal's behavior as a transgression at all. I do not disapprove of its aggressive behavior; I simply dislike it in an elaborate way.

The reason we do not make moral judgments concerning animals is because they lack a certain kind of agency that we think of as a prerequisite for moral assessment (it does not matter to our current purposes what the nature of this agency is). Taking this into account, one might respond that the emotions that form the basis of moral judgment are a kind that can be coherently deployed only toward creatures that fulfill these criteria of agency. The "dislike" felt toward a violent animal just is not the right sort of affective state to begin with (the response goes); perhaps talk of "disapproval" would be more apt than talk of "dislike."

The problem with this response is that disapproval is not a mere noncognitive response; it is a mental state permeated with conceptual content. Disapproval requires

internalization of norms requires that one "connect with these rules *emotionally*" (2012: 114) to be necessarily at odds with the cognitivist line I push in this essay.)

a concomitant judgment that the object of assessment has transgressed in a manner that warrants some sort of punitive response (if only treating with coolness). One therefore cannot appeal to disapproval as the basic noncognitive state to explain meriting (for example). The same problem would emerge if one tried to account for moral judgment in terms of the emotion of guilt—for this is an emotion with conceptually rich components (see Joyce 2006a: 101–4). I therefore doubt that one can build moral judgments out of affective phenomena alone.

Not only are purely noncognitive building blocks insufficient for moral judgment, but they appear to be unnecessary. Consider a moral judgment voiced in circumstances of emotional fatigue. Perhaps one has just been exposed to a sequence of similar moral scenarios and one's capacity for emotional arousal has ebbed. (Maybe one is ticking the hundredth box on a psychology experiment designed to ascertain subjects' moral intuitions on a range of cases.) Or perhaps one is simply distracted. All too often those who claim that emotional arousal is necessary for moral judgment focus on extreme cases: our disgust at pedophilia, our horror at the thought of the trains discharging their passengers at Auschwitz. Mundane moral judgments—like thinking that the gold medalist deserved her win, or that a person's ownership of his shoes grants him certain rights to that footwear—do not get a look in. One can claim, of course, that even for these mundane cases emotional arousal is *possible* (imagine someone having his shoes stolen; picture his outrage; visualize his suffering as he walks home barefoot through the snow), but emotional arousal to *anything* is possible.

This is one problem with Prinz's view that even if someone making a moral judgment is not emotionally aroused he or she is at least disposed to become emotionally aroused (2007: 84ff). Even if one could specify precisely what kind of emotion is relevant, there is simply no such thing as the disposition to have that emotion (occurrently) *period*; it must be a disposition to have that emotion (occurrently) in such-and-such circumstances. But while one may identify circumstances under which an individual might become emotionally aroused at the thought of someone's enjoying rights over his own shoes, so too one may think of circumstances under which an individual might become emotionally aroused at the thought that gold has atomic number 79 (or any other matter). It may be possible to find a principled distinction between such cases, but to my knowledge none has ever been articulated.

Highlighting the cognitive achievements inherent in moral judgment is not intended to exclude the affective components. As we have seen, affective mechanisms were probably central to the emergence of moral judgment—at least as pre-adaptations—and all the evidence indicates that emotions continue to play a central role in human moral life (see Haidt 2001; Greene and Haidt 2002; Wheatley and Haidt 2005; Valdesolo and DeSteno 2006; Small and Lerner 2008; Horberg et al. 2011). None of this, however, undermines the hypothesis that certain cognitive capacities are necessary for moral judgment, and that these capacities were the key development—the crucial additional building blocks—in the emergence of human morality.

The cognitive capacities I have in mind might be described as those necessary for the "moralization" of affective states. Consider the elaborate cluster of conations and meta-conations described previously, which I doubted were sufficient for a moral judgment. What the cluster seemed unable to account for were ideas like *disapproval*, *transgression*, and *merited reaction* (that is, *desert*). Without these, the fully-blown moral conceptions of *obligation*, *prohibition* (and thus *permission*[3]) are unavailable. Without the concept of *obligation*, there is no possibility of judging anyone to have a *right*, and without rights there can be no idea of *ownership* (only the idea of possession).

The chimpanzee brain lacks the mechanisms necessary to access this conceptual framework probably as surely as the human brain lacks the mechanisms for navigating the world using echo-location. Even if we could ramp up the chimpanzee's capacity for meta-conations (allowing them, say, the capacity to get angry at those who do not get angry at anyone who fails to get angry at someone who does so-and-so), we still would not thereby grant them the capability for judging a punitive response to be *deserved*. Nor would we grant them this capability if we could boost their abilities to discriminate factual data in their environment (allowing them, say, the capacity to infer that if X desires Y's welfare, and X believes that Z will get angry at Y if Y performs action ϕ, then X will want Y to refrain from ϕ-ing). It cannot be the mere "abstract" quality of moral concepts that places them beyond the chimpanzee's grasp, for in other ways chimpanzees wield abstract concepts smoothly.[4] De Waal rightly claims that humans have a greater capacity to internalize norms than other primates (Flack and de Waal 2001: 23; see also Boehm 2012: 113–31), but the puzzle remains: What mechanisms does a brain need in order to have the capacity to internalize a norm? It is natural to answer by saying something about the fear of punishment becoming assimilated, such that the individual self-regulates behavior by administering his/her own emotional punishment system. But the puzzle reiterates. To fear punishment is not to have internalized a norm (since one can fear punishment for a crime that one does not believe really is a crime); for internalization, one must believe that punishment would be *merited* and thus be disposed to dispense a kind of punitive self-reproach to oneself even in the absence of witnesses. But what accounts for this concept of "meriting"? Again I would answer that it is challenging to see how a creature could form such a thought using only purely conative and more general data-processing mechanisms (no matter how elaborate). I propose that norm internalization requires cognitive resources dedicated to normative thinking in particular.

[3] If one lacks the concepts of *obligation* and *prohibition*, then one lacks the concept of *permission*. Contra Camus' claim that "if we can assert no value whatsoever, everything is permissible" (1951), if there are no moral values then *nothing* is morally permissible.

[4] Consider a chimpanzee's postponing a vengeful act against a rival until a good opportunity arises. Perhaps we grant it deliberations about plans it will execute "later"—but *later* is an abstract concept. Or consider the way that chimpanzees can play "spot-the-odd-one-out"-type games (Garcha and Ettlinger 1979). *Sameness* and *difference* are abstract concepts.

The suggested hypothesis is that the human brain comes prepared to produce normative cognitions in a similar way that it comes prepared to encounter faces, other minds, and linguistic stimuli. This is not to say that it comes prepared for any particular normative system: that is, one with a particular normative content. The conspicuous phenomenon of moral disagreement demonstrates that moral content is learned and to some extent flexible, in the same way that the abundance of natural languages demonstrates that languages are learned and to some extent flexible. And to restate an earlier point: The hypothesis that the human brain comes prepared for normative thinking is a more general proposition than the moral nativist hypothesis. Perhaps Nichols is correct that non-hypothetical normative thinking is an adaptation while specifically moral thinking is a spin-off capacity. Or perhaps it is the other way round. Deciding whether something is an adaptation involves a large dose of inference and speculation concerning what we suppose were the relevant adaptive problems placing pressure upon our ancestors in the distant past.

Insisting on the cognitive components of moral judgment still leaves much undecided about the exact nature of these judgments. Some have argued, for example, that one characteristic of moral judgments is a particular kind of practical authority: Moral rules (unlike those of most other normative systems) are those with which one must comply whether one likes it or not. Others have doubted this, allowing that a person with sufficiently aberrant goals and desires (and appropriately situated) may well have no reason to care about moral imperatives.[5] The cognitive quality of moral judgment is consistent with either view; it is silent on the subject. A disquieting possibility is that the notion of moral judgment is in fact not as determinate on this matter (or on other matters) as we generally presuppose. Perhaps there is simply no fact of the matter as to whether moral rules have or lack this authoritative quality. Certainly people seem to generally imbue their moral prescriptions with this kind of strong authority, so maybe having a theory that provides this authority is a theoretical desideratum. But perhaps this authority is not an indispensable component of morality; maybe if we can make no sense of this authority and have to settle for a normative system lacking it, the system would still deserve the name "morality." One way of diagnosing this situation would be to say that strictly speaking morality has this authoritative quality, but loosely speaking it need not.

Something similar has been said about language by Marc Hauser, Noam Chomsky, and W. Tecumseh Fitch, who argue that one can speak of language in a broad sense or a narrow sense (2002). The former consists of linguistic capacities that we share with other animals, whereas the latter includes the uniquely human trait of linguistic recursion. There is no answer to the question of which idea captures what is "really" language; our vernacular concept of *language* is simply not so fine-grained as to

[5] Philosophers who advocate the thesis that moral prescriptions enjoy some kind of special authority include Immanuel Kant, J. L. Mackie, Michael Smith, and Christine Korsgaard. Those who allow the possibility that one may have no reason to act morally include David Hume, Philippa Foot, David Brink, and Peter Railton.

license one answer while excluding the other. Faced with the query of whether vervet monkeys, say, have a language, the only sensible answer is "In one sense yes and in one sense no."

The same may be true of morality. The vernacular notion of a moral judgment may simply be indeterminate in various respects, allowing of a variety of precisifications, with no particular one commanding acceptance. This raises the possibility that the capacity to make moral judgments construed in one sense may be an adaptation, while the capacity to make moral judgments construed in another (equally legitimate) sense is not. One might even go so far as to say that chimpanzees satisfy the criteria for making moral judgments *very loosely construed*—though I would urge against liberality taken so far. A less excessive and not implausible possibility is that on some broad construal of what a moral judgment is, the capacity to make them is a deeply entrenched part of evolved human psychology, while on a stricter construal the capacity is a recent cultural overlay: a veneer.[6]

Implications of Cognitivism

Whether on any reasonable precisification of the concept *moral judgment* the cognitive element is necessary is something on which I will not attempt to adjudicate (though earlier arguments reveal my inclination to think so). Certainly I maintain that this element is necessary at least for moral judgments strictly construed. I will close by considering some of the implications of moral judgments being cognitive in nature.

To claim that moral judgments essentially involve a cognitive component is basically to claim that they essentially involve beliefs. For example, if one holds (as one should) that a judgment that a punitive response is deserved must involve something more than just elaborate conative attitudes, then one holds that it involves (possibly *inter alia*) *the belief* that the punitive response is deserved. Once beliefs are in the picture, then certain distinctive ways of assessing moral judgments must be permitted, meaning that human morality can be interrogated in ways that, say, chimpanzee social systems cannot be. A chimpanzee group may enforce a rule that is in fact practically sub-optimal; so too may a human group. An individual chimpanzee may become affectively aroused at another in a way that harms its own interests (or furthers its own interests); so too may a human individual. But the fact that the human moral faculty involves normative beliefs means that human moral judgments can be evaluated in additional ways for which evaluating the chimpanzee response would make no sense. Beliefs can be assessed for truth or falsity in a way that purely noncognitive states cannot be. Beliefs can be assessed for justification or non-justification in a way that purely noncognitive states cannot be (this is not to claim that all talk of justification is

[6] This, clearly, would not be what de Waal means by "veneer theory," since, on the view just described, morality (strictly construed) would be a veneer over a core of social and altruistic tendencies, not (as de Waal envisages) over a core of nasty asocial selfishness.

misplaced for noncognitive attitudes, but that it must be of a very different type[7]). Therefore, a human moral response may be probed with the questions "Is it true?" and "Is it justified?" And if one can do this for a token judgment, there seems nothing to stop one posing these questions on a grand philosophical scale: inquiring of human moral judgments in general "Are they true?" and "Are they justified?"

Some may say that asking these epistemological questions of morality is somehow off the mark—that the more important question regarding human morality *is* the one that can also be asked of chimpanzee social regulation: namely, "Does it work?" I find that I have nothing to say about which kind of question is more urgent or more interesting; it's a matter of what one's theoretical concerns are. I do think, however, that the epistemological questions can be legitimately asked of any belief, and it is the job of the metaethicist to press these questions hard regarding moral beliefs.

My approach to these matters puts me somewhat at odds with that of Philip Kitcher (2011, 2014). Kitcher sees moral judgment as having emerged for a purpose, allowing one to speak of its fulfilling its function well or poorly. This in turn allows one to make sense of moral progress, but not in the manner of scientific progress—that is, the attainment of improving approximations of the truth—but in the manner of refining a tool to better accomplish its task. Moral truth, for Kitcher, can enter the picture later: defined derivatively from the account of moral progress, not vice versa.

Kitcher and I agree that a "moralization" of affective attitudes occurred at some point in our ancestry. In this essay I have advocated the view that what allowed this moralization were new building blocks of a cognitive nature: essentially, beliefs about behaviors being forbidden, punishments being just, and so forth. Instead of their large-scale cooperative projects being at the mercy of capricious conative states, our ancestors became able to think of cooperation (in certain circumstances) as *absolutely required*, of defection *meriting* penalty, and so on, which supported a more robust motivation to participate. I'm inclined to think that Kitcher is correct in holding that the purpose of morality is, broadly, to augment social cohesion, but I would place more focus on *how* moral thinking accomplishes this end: by providing people with beliefs concerning actions having moral qualities. Kitcher (2014) calls the view that moral judgments track pre-existing moral properties "a bad philosophical idea." He may be correct, and yet exploiting this "bad idea" might be exactly how ordinary human moral thinking actually functions. After all, how, one might ask, does moral thinking augment social cohesion better than altruistic sentiments? Why is the motivation to cooperate often more reliable when governed by thoughts like "It is my duty to help him"

[7] A basic distinction here is between instrumental justification and epistemic justification. Something is instrumentally justified if it furthers one's ends. Mary's belief that the famine in Africa is really not so bad may be instrumentally justified (for her) if her knowing the truth would cast her into a depression. A belief is epistemically justified if it is formed in a way that is sensitive to the evidence. Mary's belief that the famine in Africa is not so bad, though it makes her happier, is epistemically unjustified if she has been exposed to sufficient evidence of its falsehood (which she ignores). When I say that noncognitive attitudes cannot be assessed as justified or unjustified, I mean epistemic justification.

than when governed by thoughts like "Gee, I really like him"? The former, it is tempting to answer, gains motivational traction by exploiting the idea (however vaguely) of externally binding rules of conduct—imperatives that are inescapable because they do not depend upon us for their authority—moral truths to which our judgments must conform, not vice versa. (Kitcher (2014: 249) refers to the idea of a "transcendent policeman.") But such ideas will typically accomplish this social role only if they are *believed*.[8] And so long as they are beliefs, we can immediately ask "Are they true?" and "Are they (epistemically) justified?"

It is the job of the philosopher to investigate whether any sense can be made of this idea of "inescapable authority." If it cannot, then human moral beliefs may be systematically false (or, less aggressively: human moral beliefs strictly construed may be systematically false). The fact that one may nevertheless be able to speak of some moral systems serving their evolutionary function better than others—that is, to speak of moral progress—would not cut against this skepticism. To use a crude analogy: Religion may have evolved to serve some social function, and some religions may do so better than others, but for all this atheism may be true.

Conclusion

In conclusion, let me summarize what this essay has attempted via a quick clarification of two potentially misleading pieces of terminology: "naturalization" and "value."

Most of us seek a naturalization of human morality. We want to understand morality as a non-mysterious phenomenon, with a history that possibly stretches deep into our evolutionary past, accessible to empirical scrutiny. The second and third sections of this essay sought to contribute (modestly) to this goal by drawing attention to some fairly deep challenges for this program. My intention was not to scupper the project, but to face up honestly to some difficulties confronting it. But there's another kind of "naturalization" which is a whole new ball game. When metaethicists talk of "moral naturalism" they typically mean the provision of a naturalization of moral properties. This very different kind of naturalization was the concern of the fourth section.

The former kind of naturalization seeks to understand how moral judgment fits into the scientific worldview; the latter kind seeks to understand how moral goodness (and so on) fits into the scientific worldview. Obviously, one can be optimistic about the prospects of the former while highly dubious of the latter. Compare, again, the analogous two ways of understanding what it takes to provide a "naturalization of religion": One seeks to place religious practices and belief within a scientific worldview; the other would seek to locate God within a scientific worldview.

A matching ambiguity emerges when we talk of "values." It is helpful to bear in mind that "value" is both a verb and a noun. We can investigate what is going on in a human

[8] On other occasions I have explored the idea that merely *thinking* such thoughts, without believing them, might have motivational impact (Joyce 2001, 2005), but here such complications are bracketed off.

brain when its bearer values something; it is far from obvious that doing so contributes anything to our wondering what things have value. (Compare Patricia Churchland's guiding questions: "Where do values come from? How did brains come to care about others?" (2011: 12).) Of course, one might think that the latter is some sort of function of the former (in the same way that the monetary value of things depends on what pecuniary value we are collectively willing to assign them)—but this is a substantive and controversial position in moral philosophy requiring argumentative support. Many metaethicists (and folk!) think, by contrast, that "value" as a noun is the primary notion, while our valuing activity has the derivative goal of discovering and matching what values exist.

The second and third sections focused on the origins of moral valuing as an activity: worrying that it will be hard to discern whether the human trait of morally valuing things is an adaptation or a byproduct (second section), and concerned that the trait is not, in any case, well-defined (third section). The fourth section argued that if moral valuing involves beliefs (as I maintain it does), then it is always reasonable to inquire whether these beliefs are true. To do so is to focus on "moral value" as a noun—asking whether the facts that are necessary to render moral beliefs true (facts about which actions are forbidden, which are morally good, and so forth) actually obtain. Although on this occasion I have lacked the time to present any arguments, my notes of pessimism have probably been apparent.

6

The Many Moral Nativisms

John Stuart Mill's opinion that "moral feelings are not innate, but acquired" (Mill 1861: 527) was, in the estimation of Charles Darwin, destined to be judged as "a most serious blemish" on that moral philosopher's future reputation (Darwin [1879] 2004: 121). But Darwin's prophesy has so far proved incorrect; Mill's opinion on the matter has hardly been commented upon, let alone decried. Indeed, the whole question of the origin of human morality received remarkably little discussion in the century or so after Darwin's *The Descent of Man*.[1] The last two decades, however, have seen the question placed back on the agenda. The emergence of *fin de siècle* evolutionary psychology—and in particular its pioneers' decision to focus on the moralistic trait of "cheater detection" as their favorite case study (see Cosmides and Tooby 1992)—has prompted burgeoning debate about moral nativism. While this debate has yet to mature, and though one of its striking characteristics is a tendency for claims to be pressed (both for and against) with a confidence disproportionate to available evidence, we nevertheless might reasonably hope for genuine progress in the foreseeable future. Before that progress can occur, however, we need to understand the hypothesis. Currently there are a number of points of significant imprecision in the debate over moral nativism that often pass unnoticed and which lead to seemingly opposed factions speaking at crossed purposes. I think it is fair to say that we are at present in the same state that William Darwin (in a letter to his father) attributed to Mill: of being "rather in a muddle on the whole subject."[2]

In previous works I have advocated moral nativism (Joyce 2006a, 2006b)—though I did so provisionally and cautiously; my objective was concerned more with clarification than all-out endorsement. Advocating moral nativism is not my intention in this essay; my goal here is principally diagnostic. I will highlight three places where the

[1] Of course, one would have little trouble assembling a list of books and articles from 1880 to 1980 (say) that would appear to counter this claim (Edvard Westermarck's works in particular spring to mind); but I would maintain that this list—though superficially impressive if gathered in an endnote—still constitutes "remarkably little attention" for a century's-worth of intellectual labor on the topic.

[2] Darwin Archives: DAR88.76–7. Charles had evidently asked William to read and summarize Mill's *Utilitarianism* for him while he (Charles) was preparing the second edition of *The Descent of Man*. Given that the point of this delegation of labor was to discern Mill's views on the origin of the moral sense, I cannot resist remarking that it was William who had many years earlier been the subject of his father's article "A biographical sketch of an infant," and whose "first sign of moral sense" was observed at just over a year old (Darwin 1877: 291).

nativist/non-nativist debate fragments in such a way that it ceases to be clear what the hypothesis is that is under dispute. In two of the three problematic places, the options for reinstating precision are reasonably well defined, so my conclusion is that disputants simply need to take care to specify which understanding of the hypothesis is under discussion. In the third case, however, my attitude is rather more pessimistic. Here, it seems, we find at the heart of the debate an inchoate concept—that of *moral judgment*—regarding which the options for precisification are not well understood, and for which any stipulative specificity appears more of a misleading distortion than a welcome clarification. One possible consequence of this is that on some legitimate conceptions of *moral judgment* moral nativism is true, but on other equally legitimate conceptions moral nativism is false. And if there is no satisfactory way of deciding among these conceptions, then the debate over moral nativism is undecidable—not just in the sense that we lack decisive data, but in the sense that there is really no fact of the matter.

The First Node of Imprecision: Innateness

In its crudest form, moral nativism is the view that human morality is innate. What might be meant by "human morality" is a question that will occupy much of this essay, but first our attention should pause on what is meant by "innate." Some participants in the debate over moral nativism know what they mean by "innate," but many employ an intuitive folk notion that doesn't withstand critical scrutiny. Of those that do have a clear view of what they mean, not all mean the same thing.

The folk notion of innateness is a blend of several subclusters of ideas. One such group of ideas pertains to a trait's being present at birth, to its being not learned, to its being determined by genes rather than environment, to its being developmentally robust in the face of environmental variation. Another idea central to innateness is the Darwinian notion of a trait's existing because it was selected for by the process of natural selection—that is, of a trait's being an adaptation. Another is the essentialist idea of a trait's being species-typical: present in all members of the species or at least in all "normal" members. (For diagnosis and discussion of such options, see Griffiths 2002; Mameli and Bateson 2007; Mameli 2008.)

These ideas are not all equally scientifically respectable, and, more to the point, they are far from coextensional. Down's syndrome is present at birth, genetically influenced, and developmentally robust, but it is not an adaptation. The possession of a certain stone-knapping technique may satisfy the criteria for being an adaptation (it may be transmitted from parent to offspring and may owe its existence to the fact that it enhanced reproductive fitness), but is neither non-learned nor developmentally robust. And so forth. Hence we must reject the common unexamined presupposition that these phenomena more or less come together and thus can be treated as facets of a single "cluster" concept. In light of the way these disparate ideas get lumped together, Matteo Mameli (2008) disparagingly refers to *innateness* as a "clutter" concept.

In the literature on moral nativism, two conceptions of innateness are most conspicuous: an evolutionary conception and a developmental conception. A typical statement of the evolutionary conception comes from Jesse Prinz, who sums up moral nativism as the claim that "morality is an evolved capacity" (Prinz 2009: 168). I have myself described moral nativism as the view that "morality (under some specification) . . . is to be explained by reference to a genotype having granted ancestors reproductive advantage" (Joyce 2006a: 2). On this view, moral nativism is the claim that morality is a Darwinian adaptation.

Standing in contrast to this is the developmental conception, according to which the emergence of the trait is buffered against variation in the developmental environment (Ariew 1996, 1999). Chandra Sripada and Stephen Stich use such a conception when they write that "we can consider a normative rule to be innate if various genetic and developmental factors make it the case that the rule would emerge . . . in a wide range of environmental conditions" (Sripada and Stich 2006: 299).

These two conceptions of innateness are by no means coextensional. The trait of morality might be a specific adaptation but may nonetheless require particular structured environmental inputs in order to become manifest. If such inputs were reliably available in the environment in which morality evolved, then there would be no selective pressure to make the developmental emergence of morality robust in the face of environmental perturbation. The reverse is also true: Human morality may be developmentally canalized while not being an adaptation. Some of the well-known ways by which traits may become canalized without being adaptations—genetic drift, mutation, genetic disease—are admittedly far-fetched in the case of morality. However, one way is entirely plausible: that morality is a byproduct of other adaptations. (This possibility will be discussed later.)

Clearly, this introduces potential confusion into the debate over moral nativism, for it allows that moral nativism may be true in one respect but false in another. Even when advocates of a particular view are conscientious in articulating which thesis they mean to defend or attack, casual readers may miss the qualification. For example, in his paper "Moral nativism: A sceptical response," Kim Sterelny is careful to explain that he is skeptical of the *developmental* nativist thesis. He allows that "there is a plausible . . . case for the idea that moral cognition is an adaptation," but adds that "even if that is right, it does not follow that this capacity is innate" (Sterelny 2010: 280). If such comments are overlooked, however, then one might gain the impression that Sterelny is in the same camp as other opponents of moral nativism when in fact these others are skeptical of the *adaptational* nativist thesis. More worryingly, one may gain the erroneous impression that Sterelny is in the opposing camp to someone like myself, who has advocated the adaptational nativist thesis, when it is entirely possible that we agree on everything of substance.

We have seen that in assessing the thesis of moral nativism possible misunderstandings lurk around the term "innate." Yet the possible misunderstandings surrounding the term referring to the trait in question—"human morality"—are even greater. In

subsequent sections I will tease this matter apart into two further particular points of imprecision, but first I will introduce the general problem via a discussion of altruism and Darwin's views on moral nativism. My reason for doing so is as follows. I want to demonstrate that the trait in question, human morality, is difficult to define with any precision—that it admits of more liberal and more strict characterizations. To illustrate this, it is useful to begin with another trait, altruism, that is not a million miles from morality but which is pretty clearly not the same thing. Identifying the difference between altruism and morality forces us to ask what exactly the trait of "morality" is. Darwin's own views are worth discussing here because he begins with prosocial attitudes, like altruism, which he then supplements with further psychological traits in order to achieve something that, he believes, deserves the label "the moral sense." (Darwin, obviously, is focused on evolutionary rather than developmental emergence.) This transition from non-moral organism to moral organism is exactly what we are interested in. But Darwin's efforts also exemplify the difficulty and obscurity of the task—the fact that it is radically unclear what an adequate account of the transition from the non-moral to the moral would have to involve. There is, I wish ultimately to argue, no single answer to this question.

From Altruism to Darwin

It is standard to distinguish two forms of altruism: psychological and evolutionary. An action is psychologically altruistic if and only if it is motivated by an ultimate desire for the well-being of some other organism. A behavioral trait is evolutionarily altruistic if and only if it benefits another at some cost to the individual, where benefits and costs are understood in terms of reproductive fitness. (It must be added that the trait has been selected *because* it benefits another, otherwise one ends up counting as altruistic such things as a sea turtle's drive to lay its eggs on the beach, which makes its hatchlings such easy prey for seagulls.) The former is an articulation of a vernacular notion, whereas the latter is very much a term of art.

The extensive literature ostensibly concerning the "evolution of altruism" often fudges this important distinction, and, indeed, frequently concerns neither. Consider the so-called altruistic behavior of bees. It is surely not psychologically altruistic (since bees simply lack the motivational prerequisites), but nor is it obviously evolutionarily altruistic: William Hamilton's breakthrough work on kin selection (Hamilton 1964) demonstrated how the individual bee who dies to save her nest-mates is in fact advancing her own inclusive fitness. Or consider the reciprocal grooming behavior of primates (see Schino and Aureli 2010). If the explanation of primate A's tendency to take the time and effort to groom primate B is that this increases the probability of A's being groomed in return, then in performing this behavior A is reproductively better off (eventually) than if it did not. (See West, Griffin, and Gardner 2007.) Similar considerations pertaining to hunting lions, mobbing birds, meerkats on sentry duty, and so on, will also reveal neither psychological nor evolutionary altruism. For this reason,

it is best to call such behaviors simply "cooperation" (leaving this an intuitive term), which then allows the questions of whether these cooperative behaviors are also instances of psychological altruism or evolutionary altruism to be substantive inquiries. (For discussion of how true evolutionary altruism is possible, see Sober 1988; Sober and Wilson 1998.)

Without pausing to investigate the details of how much cooperation in nature really is evolutionarily altruistic, one can at least safely say that cooperation often turns out to be evolutionarily selfish, in the sense that the cooperative behavior ultimately enhances the actor's reproductive fitness better than not cooperating. The temptation that it is crucial to resist is thinking that this evolutionary selfishness has any bearing on psychological selfishness. Organisms that do not have psychological states at all, such as plants, may be evolutionarily selfish or altruistic. To satisfy the prerequisites for being *psychologically* altruistic or selfish, a creature must be able to have ultimate motives concerning others' or their own welfare, which requires them to have the concepts of *other* and *self*. The only creatures for which we can be confident of the satisfaction of these prerequisites are humans.

I will take it as obvious that the mere fact that a behavioral trait is to be explained by reference to *evolutionary* altruism is insufficient to make the introduction of talk of "morality" appropriate. A plant may have evolutionarily altruistic traits, but the plant neither makes moral judgments nor is a suitable subject of our moral appraisals. But it is not so obvious that there is no connection between *psychological* altruism and morality, so this requires some discussion. It is particularly important here because a plausible case can be made that psychological altruism in humans is innate; hence this may have direct implications for the prospects of moral nativism.

The details of the argument for nativism concerning psychological altruism need not delay us; a sketch will suffice for present purposes. The argument concerns *evolutionary* nativism rather than developmental nativism, and has been advocated by Elliott Sober (2000). Sober's principal opponent is the psychological egoist, who holds that all human actions are performed with the ultimate motive of benefiting the actor. Given that natural selection has clearly forged humans to be cooperative in certain ways—at the very least, caring for our offspring—Sober wonders what kinds of psychological mechanisms would likely be favored to govern these cooperative tendencies. Assuming that it is adaptive to come to the aid of one's children when they are in distress, for example, what is the better psychological setup? On the one hand, we can envisage a parent motivated to provide aid simply because he loves his daughter—he cares directly for her in such a way that a perceived threat to her welfare directly prompts action. On the other hand, we can imagine the egoistic parent: moved via a combination of the belief that his daughter's suffering has a negative effect on his own welfare plus his love for himself. One might plausibly claim that the former mechanism is more reliable and less complicated—and thus, *ceteris paribus*, more adaptive—than the latter. By analogy, a person prompted to withdraw her fingers from a flame *by pain* seems moved by a more reliable and less complicated process than a person who forms

a belief about the bodily damage caused by fire and calculates the costs and benefits of action versus inaction. This argument may not be without problems (see Stich 2007), but here my intention is not to evaluate or endorse the argument, but rather to examine what would follow—or, more precisely, what would *not* follow—if it were sound.

We have seen that to be psychologically altruistic a creature needs to be fairly cognitively sophisticated, but it doesn't follow that the creature is therefore capable of making moral judgments. This truism is potentially muddied by the fact that the only clear-cut case of a species capable of psychological altruism (and selfishness) is also the only clear-cut case of a species of which we speak in moral terms: namely, humans. Still, the conceptual distinction does not seem difficult to discern. One can imagine members of a cognitively sophisticated social species, motivated by love and altruistic tendencies toward their fellows, but who fail to "moralize" these feelings—who are, in fact, constitutionally incapable of making a moral judgment. Such creatures have powerful desires to see their loved ones flourish, but cannot conceive of actions satisfying those desires as morally right or obligatory.

It might be conceded that these imaginary creatures don't make moral judgments but maintained that they are at least morally *praiseworthy* (that is, that they warrant *our* moral judgment). But upon reflection even this is unclear. After all, altruistic motives can prompt someone to act in a morally despicable manner. Consider a mother who genuinely adores her child, and who poisons all the other children at the sports day so her child can win. In any case, it seems misguided to identify moral nativism with the claim that the trait of *being morally praiseworthy* is innate. Such a proposal would lead straight into a metaethical quagmire from which the debate is unlikely ever to emerge. We are not primarily interested in the question of at what point, either in evolution or development, humans become *morally admirable*; we are interested in at what point they become capable of *making moral judgments*. Popular discussions of moral nativism with headlines like "Are we born to be good?" or "The moral animal" or "Chimps display morality" (and so on) blur this basic distinction, and in doing so spread more misunderstanding than illumination.

Once we focus nativism on the question of *moral judgment*, it becomes clear that we are asking about something different from (or perhaps more than) psychological altruism. None of this is to deny that the emergence of psychological altruism (both evolutionarily and developmentally) might be a crucial precursor to moral judgment; I'm not claiming that someone with an interest in moral nativism should dismiss all discussion of the emergence of psychological altruism as irrelevant. My claim is simply that moral judgment is not the same thing as altruism, and that establishing nativism about altruism does not establish moral nativism. Although this much seems assured, the natural further question of what exactly *is* required for moral judgment is much harder to answer.

As a way of illustrating this problem, I turn now to Darwin's views on the matter. Darwin undertakes the task of supplementing prosocial emotions (like altruism) with further psychological capacities in an attempt to "build" a human moral sense. Drawing

attention to the difficulties inherent in this project is one of the goals of this essay, so sketching his attempt is a useful exercise.

Darwin is no psychological egoist. He writes:

> With respect to the impulse which leads certain animals to associate together, and to aid one another in many ways, we may infer that in most cases they are impelled by the same sense of satisfaction or pleasure which they experience in performing other instinctive actions . . . In many instances, however, it is probable that instincts are persistently followed from the mere force of inheritance, without the stimulus of either pleasure or pain . . . Hence the common assumption that men must be impelled to every action by experiencing some pleasure or pain may be erroneous. (Darwin [1879] 2004: 128)

Darwin speaks frequently of the "social instincts" of animals—which include such affections as sympathy, love, and pleasure in the company of one's fellows—and there is no doubt that he considers these capacities to be psychological adaptations. But he is equally adamant that they do not suffice for a moral sense: "I fully subscribe to the judgment of those writers who maintain that of all the differences between man and the lower animals, the moral sense or conscience is by far the most important." He goes on:

> [A]ny animal whatever, endowed with well-marked social instincts . . . would inevitably acquire a moral sense or conscience, as soon as its intellectual powers had become as well, or nearly as well developed, as in man. (Darwin [1879] 2004: 120–1)[3]

What are these "intellectual powers"? First of all, Darwin thinks, one needs a good memory, in order to recall those times in the past when one has failed to act cooperatively and (as a result of one's social instincts) felt dissatisfaction. One needs to recall that the benefits gained from failing to cooperate (that is, the profits of defection) were fleeting. Second, the emergence of language allows that "the wishes of the community could be expressed, [and] the common opinion how each member ought to act for the public good, would naturally become . . . the guide to action" ([1879] 2004: 122). Lastly, one needs the capacity to form *habits* of acting for the good of one's fellows.

This might be interpreted as an argument for moral nativism (of the adaptational variety), but on another interpretation Darwin thinks of the moral sense as a kind of "spandrel" derived from faculties that evolved for other purposes.[4] In fact, he is explicitly undecided on the matter. Referring just to the social instincts, he writes that it is "impossible to decide in many cases whether certain social instincts have been acquired through natural selection, or are the indirect result of other instincts and

[3] Darwin uses "moral sense" and "conscience" seemingly interchangeably. One interesting implication is that he sees the moral sense primarily in terms of *self-directed* moral evaluations—for that is what a conscience is. It seems to me, moreover, that this gives license to assume that when Darwin talks of a "moral sense" it is a faculty of making moral *judgments* that is under discussion. While I am aware that there is some room for debate about this assumption, here I'm willing to forgo argument and treat it as a simplifying supposition.

[4] In previous work I have interpreted Darwin as a moral nativist; I now think that this is not straightforward.

faculties" ([1879] 2004: 130). We will return to this distinction later; currently what interests me is how, precisely, the moral sense is supposed to emerge from these elements. My strategy will be to grant Darwin these ingredients and attempt to motivate doubt that we have enough to warrant the label "a moral sense."

Darwin certainly has plenty of persuasive things to say about the evolution of the social instincts; on this topic he is squarely in his "comfort zone." But his explanation of how certain "intellectual powers" get married to those instincts, resulting in a moral sense, is considerably sketchier and less convincing. The latter two ingredients listed in his initial presentation—language and habit—hardly get a further mention. (It is, besides, unclear whether he thinks of these two traits as *necessary* for a moral sense.) It is the role of *memory* that he mentions repeatedly and evidently judges of paramount importance. But the case is underdescribed at best.

Consider a creature brimming with altruistic sentiment for its conspecifics. I argued above that this alone does not suffice for a moral judgment. The creature doesn't think that it *ought* to help its fellows; it doesn't think of failure to help as *prohibited*; it doesn't think that such failures warrant punishment or disapproval, or that helping merits praise. It simply *wants* to help. Yet suppose that occasionally the creature experiences temptations to do otherwise, since there are other competing instincts operative in its psyche. When this creature succumbs to such temptations, it enjoys the satisfaction of the tempting outcome (whatever it may be) and yet also feels bad because of the frustration of its natural desire to cooperate. Let us stipulate that the creature's instincts are such that the pleasures achieved at the expense of cooperation tend to be short lived. Let us now grant it the intellectual powers both to realize and to remember this fact. Thus, when temptation arises, the creature is able to deliberate along the lines of: "Well, that sure looks enticing, but I remember how rotten I felt last time I succumbed to temptation, so I'll cooperate." So now we have a creature with self-control in favor of cooperative behavior.

But where does the *moral judgment* emerge in this process? Acting cooperatively is still, essentially, just something that this creature *wants* to do. Compare a monkey that is often tempted to climb its favorite tree using the dangling outer branches, but who, through trial and error, comes to learn that it is safer to ascend by the trunk. When faced with the temptation to dart up the dangly branches the monkey may pause and recall the bruises of earlier decisions. So now we have a creature with self-control in favor of climbing a tree via the trunk. But do we credit the monkey with anything like the judgment that climbing the outer branches is a *transgression*? If it does climb by the outer branches without mishap, we can imagine it thinking "That was a bit stupid, but, phew, I got away with it!" Where would be the *guilt*? Where would be the thought that it *deserves punishment* for its crime? Why would it take an interest in punishing other monkeys that exhibit foolish climbing habits?

In the case of the first creature whose instincts are in favor of cooperation, we need also to factor in the reactions of its conspecifics, but I don't believe that this alleviates the puzzle. The conspecifics *don't like it* when the individual defects on some cooperative

enterprise, and we can imagine that their disappointment and anger is something that our individual will take into account. It controls itself by remembering how bad its failures to cooperate made it feel in the past, and when those failures are accompanied by its fellows expressing their anger with (say) violence and ostracization, then self-control will be all the easier since the negative repercussions of such failures will be even worse. Thus the influence of the conspecifics will certainly significantly strengthen the process of self-control, but it in no obvious way brings about a change in *kind* in the sorts of judgments and attitudes that we attribute to the individual.

It appears, therefore, that one can identify elements that seem important to moral judgment—such as the ideas of transgression, guilt, and desert—for which Darwin's hypothesis does not account. In assessing this matter one needs to be wary of projecting one's own "moralizing" thoughts onto the imaginary characters involved. It is difficult to cleanly imagine someone simply *not wanting* to perform non-cooperative actions (in part because she recognizes that other parties *don't want* her to) without positing the seemingly innocuous extra assumption that she also judges that she *ought not* perform those actions. It is natural for us to assume that as our imaginary creature forms the habit of acting cooperatively, surely at some point it "internalizes the norm": Its *expectation* of negative outcomes morphs naturally into the thought that such outcomes are *warranted*; its *desire* for its fellows' welfare gradually begets the judgment that acting for their welfare is *desirable*; it moves from habitually not wanting something to judging it prohibited; and so forth. But assuming that this transition occurs naturally is exactly what we must not do in this context, for how such a transition occurs is precisely what is under scrutiny.

Darwin brings the discussion to the edge of "moralization," but it is not obvious that he succeeds in crossing the conceptual gap. Perhaps the ingredients he provides suffice for a thin notion of moral judgment, but there is a richer folk conception whose evolutionary emergence remains mysterious. As we shall see, the same can be said of some modern participants in this debate: They provide ingredients that may be adequate to account for moral judgment in some attenuated sense but which fail to explain important components of a robust conception of moral judgment. Thus the debate founders not merely through lack of empirical data, but through an absence of any single phenomenon uniquely deserving of the name "moral judgment." Before discussing this matter further, however, I should like to note another source of confusion about the nature of the trait whose origin is under discussion.

The Second Node of Imprecision: Content versus Concept

It is important to distinguish between moral concepts and moral judgments. Let us say that a complete paradigm moral judgment consists of the application of a moral concept, like *moral wrongness*, to a general subject, like *incest*, or to a particular subject,

like *John and Mary's incestuous relation*.[5] Given this framework, we can identify another way in which moral nativist hypotheses may vary.

One version of moral nativism will allow that certain complete moral judgments are innate. There is certainly nothing to be said in favor of the claim that complete moral judgments concerning *particulars* are innate. For example, to hold that the judgment "John and Mary's incestuous relation is morally wrong" is an adaptation would involve accepting that our ancestors somehow knew about the individuals John and Mary and formed a moral opinion about what they got up to, and that this opinion enhanced reproductive fitness. Given that a great many of our moral judgments do concern particulars, nativism about complete judgments is going to be utterly implausible for a great many of our moral judgments. Even for those moral judgments that take universals as subjects, nativism concerning the complete moral judgment is feasible only when the subject is something that was present in the environment of evolutionary adaptiveness (the EEA). One may, for example, countenance nativism for "Incest is wrong," but nativism for "Shoplifting is wrong" is a non-starter.

Another version of nativism eschews any commitment to complete judgments being innate and prefers the image of a moral faculty as a "toolkit" of moral *concepts*, with the individual's socialization process as the sole determinant of to which subjects these concepts get attached. Thus, according to this hypothesis (expressed in simplistic terms), a concept like *moral wrongness* is innate, and one social environment may lead the individual to apply the concept to *incest*, another environment may lead the child to apply it to *John and Mary's incestuous relationship* but not to *Ptolemy and Cleopatra's incestuous relationship*, while yet another may lead the child not to apply the concept to any incestuous relationship.

These two nativist positions represent extremes, between which lie a variety of hypotheses. Some allow that a few broad abstract moral principles are innate but that the environment sets the parameters of how these create specific moral judgments (Hauser 2006). Some allow that content is learned but that the moral sense comes

[5] This statement may seem metaethically question-begging and also surprising in light of other claims I have just made, so a couple of quick explanations are called for. First, at this stage of the discussion I don't intend this notion of "applying a concept" to be theoretically deep; thus my claim is meant to be metaethically neutral. I take it that the locution "applying a concept" is something that even the modern noncognitivist will seek to accommodate. Simon Blackburn's quasi-realist program sets out to "earn the right" to such realist-sounding talk but from an anti-realist position that eschews any genuine metaphysical commitment to such entities. (See Blackburn 1993, 1998.) Second, given the emerging worries about the indeterminacy surrounding the notion of *moral judgment*, one may wonder on what grounds I can confidently make such an assertion. The answer is that even if there are thinner and richer explications available of the notion of *moral judgment*—such that the former counts certain things as moral judgments that the latter will not—nevertheless, there is surely a class of paradigm instances of moral judgments to which all parties will agree. Of these paradigms, though disagreement may remain concerning in virtue of what they count as moral judgments, it hardly follows that we can say nothing about their characteristics. The statement to which this note is appended is intended to be just such a platitudinous description.

"prepared" to latch on to certain domains more easily than others (see Haidt and Joseph 2004; Sripada 2008).[6]

Even with some options in the moral nativism spectrum sketched in so heavy-handed a manner, we have seen enough to recognize that evidence favoring one version of moral nativism will not favor another. As a way of illustrating the muddle that ensues, I will examine the debate over moral universals.

In fact, even if all parties were in complete concurrence regarding which trait is under scrutiny, the place of universals in the debate over nativism would be far from straightforward. The tempting assumption that if a trait is innate then we can expect to find it manifest everywhere must be rejected. If one is focused on developmental innateness, then many innate traits are not universal (such as Down's syndrome, eye color, lactose tolerance). If one is discussing adaptational innateness, then innate traits may well require substantial environmental input—input that may have been reliably present in the EEA but is absent, patchy, or distorted in the modern environment. I intend to put these important complications aside, however, in order to focus on another simpler point about universals. For the sake of argument let us allow the assumption that innate traits will reliably emerge and thus tend toward universality. The question is: For what kind of universals should we be looking? And the answer is: It depends which version of moral nativism is under scrutiny.

In one of a series of papers arguing against moral nativism, Prinz discusses three possible moral universals: *don't harm innocent people*; *respect and obey authorities*; and *incest is prohibited* (Prinz 2009; see also Prinz 2008a, 2008b, 2014). He carefully examines historical and anthropological evidence in an attempt to find counterexamples to the claim of universality for each, thus discrediting moral nativism. But the limitations of this strategy should by now be clear: Many moral nativisms will not hold that such complete moral judgments are innate.[7]

This is not to say that Prinz's efforts are wasted. Certain versions of moral nativism may well claim that precisely these three complete moral judgments are innate, and I share Prinz's determination to reject such views. Prinz, moreover, knows that he is challenging only one form of moral nativism. He is aware of the kind of toolkit moral nativism mentioned earlier—which holds no complete moral judgment to be innate but rather postulates innate moral concepts. Prinz labels this kind of moral nativism "minimal" (Prinz 2009) and "weak" (Prinz 2014). I confess to finding this labeling system unfortunate, since it allows the anti-nativist to proceed by first refuting the "strong" versions of moral nativism (the kinds that were never terribly plausible in the first

[6] Note my avoidance of speaking of "innate moral *knowledge*"—an unnecessary practice that seems to beg several large questions. Moral nativists who seemingly lack such qualms include Sue Dwyer (Dwyer 2009; Dwyer, Huebner, and Hauser 2010) and John Mikhail (2008).

[7] Another potential problem is that Prinz sets out to investigate the existence of *cultural* universals, whereas if nativism did imply universality, we should be examining evidence of *psychological* (individualistic) universals. For the sake of argument I'll play along with the focus on cultural universals. See Buller (2006: 457–8) for critical discussion.

place), thus giving the impression of the moral nativist retreating to an ever weaker position in a desperate bid to defend his or her hypothesis. The rhetorical narrative this suggests is inaccurate, and is exasperating to anyone who begins with a desire to defend toolkit moral nativism while agreeing wholeheartedly that there is little to be said in favor of the more content-complete versions of nativism.

If we are investigating evidence for and against universality, and have the more plausible toolkit kind of moral nativism in mind, then we should be examining whether any cultures lack moral judgments altogether. If one culture thinks that incest is morally acceptable while another judges it repugnant, this is no counterexample to universality, for both cultures are still evaluating the world in moral terms. All too often the debate has revolved around the question of whether "moral universals" exist, but if I am correct then this is misguided; what we should be investigating is whether *having a system of moral judgments* is a human universal. And while it is not my intention on this occasion to press the case in favor of this latter hypothesis, it is reasonable to suppose that the prospects of its being true are far better than the likelihood of finding moral universals.

Prinz is certainly unable to provide a counterexample. At one point, he mentions the Ik group of Uganda, famously described by anthropologist Colin Turnbull (1972) as a "vicious people" with "sadistic customs." We now know that Turnbull's account of the Ik was flawed in numerous ways (see Heine 1985; Knight 1994), but even if that were not so, the "viciousness" of which he spoke is compatible with the Ik having a moral system—one that might seem blighted and alien to us, but a moral system nonetheless. Indeed, when, several years later, the Ik elders heard of how Turnbull had portrayed them to the world, they were angry that he had "spoilt" their reputation, and threatened to make him "eat his own faeces" if he ever showed his face again (Heine 1985: 3). To the extent that they thought that Turnbull *deserved* this unenviable fate, the Ik proved themselves capable of wielding a moral concept.

Prinz doesn't seriously think that the Ik lack any moral system. When he squarely addresses the toolkit version of moral nativism, he admits "I certainly don't know of any exceptions to this claim" (Prinz 2008a: 386). This concession forces a change of tactic in his pursuit of the non-nativist agenda: He moves from trying to provide counterexamples to universality and instead sets out to demonstrate that an appeal to nativism is not required to explain moral judgment; he endeavors to provide an empiricist explanation of the (possibly universal) phenomenon. In doing so, he aims to discredit a focal argument in favor of moral nativism: the poverty of the stimulus (POS) argument. According to this argument, the capacities evident in moral cognition are acquired in a manner that far outstrips the information that is available in the learning environment. The structure of the argument comes, of course, from the debate over nativist explanations of human linguistic abilities (see Chomsky 1967, 1990), where the POS argument is widely judged to be triumphant in establishing some form of nativism.[8] It is not my intention here to evaluate the prospects of a moral

[8] For powerful criticism of this orthodoxy, see Cowie (1999).

POS argument, but rather point out how progress gets confounded by distinct theoretical options being conflated.

One obvious way of countering a POS argument is to show that the stimulus is in fact a great deal less impoverished than one might have thought. Thus moral non-nativists are eager to point out how rich is the moral learning environment of the child. Shaun Nichols reminds us that "the child is exposed to *lots* of admonitions and instruction in the normative domain. Parents and teachers are constantly telling kids what shouldn't be done" (Nichols 2005: 358). Sterelny makes a similar observation:

> The narrative life of a community—the stock of stories, songs, myths and tales to which children are exposed—is full of information about the actions to be admired and to be deplored. Young children's stories include many moral fables: stories of virtue, of right action and motivation rewarded; of vice punished. So their narrative world is richly populated with moral examples. (Sterelny 2010: 289)

This is all undeniable. The child's moral world is richly structured, and the explicit moral instruction is coordinated and unrelenting.

It is not sufficient, however, simply to remark upon the wealth of the moral stimulus in a general way. We need to decide which version of moral nativism is under discussion, for this determines what kind of moral task it is whose acquisition process is under scrutiny. If our interest is in toolkit moral nativism, then focusing on how children acquire complete moral judgments is misleading; rather, our attention should be on how children acquire their basic moral conceptual tools. If this is the target trait, then wondering how children acquire the belief that shoplifting is wrong (say) would be a distraction (for I'm sure all parties can agree that they are taught it by adults); instead we should be wondering about how children acquire the concept of *moral wrongness* in the first place. Is the environment rich enough to provide them with *that*?

This is a crucial disambiguation to make before assessing the prospects of any moral POS argument, yet it still leaves progress hampered by a serious conceptual imprecision, for one is still left wondering "What is a moral judgment?" The possibility remains that moral nativism may be more plausible with certain conceptions than others. This is discernible in anti-nativist attempts to oppose the moral POS argument, as the following short review will demonstrate.

Some Anti-Nativist Hypotheses

The opponent of moral nativism will usually try to account for the human trait of making moral judgments by calling attention to other psychological traits that evolved or develop for other purposes. Often moral judgment is described as a byproduct or "spandrel" of these other traits. I will sketch a few anti-nativist views in order to give a flavor of the approach.

Prinz attempts to account for the evolutionary emergence of moral judgment from a cluster of other evolved faculties, each of which has a more general role. At the center

of his argument is the view that moral judgments are emotional responses.[9] In one paper (Prinz 2009) he proposes to construct a moral response out of emotions that are not distinctively moral: anger and sadness. We feel sad in many circumstances, but when we feel sad at having transgressed against a norm, Prinz argues, then the sadness is called "guilt." "Guilt is an accidental byproduct of sadness" (2009: 183). In other works, Prinz develops a somewhat different empiricist hypothesis. In his (2008a) he mentions not only the non-moral emotions, but some additional traits: *meta-emotions* (emotions directed at our own emotions or at others' emotions), *perspective taking* (allowing for third-party concern), and other *non-moral preferences* (such as the "social instincts" which were Darwin's starting point). The important point is that these are all general cognitive skills; thus, if moral judgment is a natural byproduct of these traits, the moral nativist would be defeated.

The view that emotion has a central role in moral judgment is also at the heart of Nichols's attempt to provide an empiricist account of the origin of moral judgment (Nichols 2005). Nichols allows that "rule nativism" might be reasonable, where the rules in question are non-hypothetical. "There is no obvious story about how the empiricist learner might come to acknowledge *non-hypothetical imperatives*" (2005: 357). He correctly argues that morality is but a proper subset of non-hypothetical rule systems, citing etiquette and institutional rules (for example, of a gentlemen's club) as involving non-moral but non-hypothetical imperatives (following Philippa Foot 1972). A key question, then, is what is distinctive about *moral* non-hypothetical imperatives. Nichols's answer starts by noting the distinctive subject matter of morality: namely, that it pertains to harm.[10] Given this characterization of morality, the second ingredient in Nichols's hypothesis is an innate affective mechanism that responds to suffering in others. This emotional response imbues a certain subset of non-hypothetical imperatives with a particular flavor (call it "moral"), picking them out as salient, resonant, and memorable. Nichols concludes:

> [B]oth of the mechanisms that I've suggested contribute to moral judgment might well be adaptations. However, it is distinctly less plausible that the capacity for core moral judgment itself is an adaptation. It's more likely that core moral judgment emerges as a kind of byproduct of (*inter alia*) the innate affective and innate rule comprehension mechanisms. (Nichols 2005: 369)

Another anti-nativist argument comes from Sterelny, though, as noted earlier, he is focused more on the developmental trajectory than the evolutionary emergence of the

[9] Or so Prinz claims when he's summarizing his view, but the more detailed presentation is rather more complicated. First, it turns out that having emotions is just the "standard" way to assess things morally (Prinz 2007: 42). Second, moral judgments are linked by Prinz not directly to emotions but to *sentiments*—where a sentiment is a disposition to have an emotion (2007: 84). Thus Prinz has at least two "escape routes" should evidence come forward of moral judgments made with no emotional arousal. For further criticism of Prinz's view, see Joyce (2009b).

[10] Nichols is aware of moral norms that have nothing obvious to do with harm (concerning, for example, cleaning the toilet with the national flag), but he states that "it is plausible that judgments about harm-based violations constitute an important core of moral judgment" (Nichols 2004: 7).

trait. Like both Prinz and Nichols, Sterelny holds that one of the key psychological ingredients in a non-nativist explanation of moral judgment is emotion. He appears willing to endorse nativist hypotheses for emotional contagion, for sensitivity to interactions involving harm, and for the emotions associated with "reciprocation, sympathy, empathy, disgust, and esteem" (Sterelny 2010: 293). He argues at length that moral learning is largely a matter of generalizing from exemplars—which explains why moral intuitions can be fast and automatic—and also stresses that this would not mark moral learning as unusual (that is, the faculties involved in prototype-comparison learning are general mechanisms). Sterelny further persuasively emphasizes the extremely rich and structured nature of the moral learning environment, arguing that the "parental generation engineers the informational environment in which the next generation develops, thus guaranteeing the development of moral competence" (2010: 294). Sterelny concludes that moral norms "are grafted on top of our dispositions to respond emotionally" (292), that moral cognition "is a natural development of our existing emotional, intellectual and social repertoire" (293), and that moral cognition "develops from an interaction between emotions, exemplar-guided intuitions and explicit principles" (293).

Clearly, it is beyond the ambitions of this essay to attempt to analyze or refute these proposals in detail; I aim to make a more general point. First, I will pursue the same strategy as was deployed earlier against Darwin: taking the ingredients offered and questioning whether they suffice for making a moral judgment. My ultimate goal, however, is not to declare that all such arguments simply fail, but rather to argue that different conceptions of moral judgment are in play.

Consider, first, Prinz's argument that guilt is just sadness directed at having transgressed against a norm. There appear to be important components of full-blooded guilt that remain unaccounted for. Sadness predicts social withdrawal, whereas guilt (unlike shame) urges reparative action (Tangney and Fischer 1995; Tangney et al. 2013). Extreme sadness cripples a person's capacity to engage in everyday activities, whereas guilt, even acute guilt, is a burden that a person can usually shoulder while getting on with things. Even the manifestation of weeping that we associate with sadness we do not associate so readily with guilt (which is not to deny that guilt can cause a person to cry[11]). Indeed, language itself should be a giveaway here. We do have words for some special instances of sadness defined according to their object. "Grief," for example, denotes sadness directed at the loss of someone or something dear to us. Notice that just as we can say "I feel grief about Fred's death," we can say "I feel really sad about Fred's death," and no one will bat an eye-lid. But compare the huge difference between saying "I feel guilty about having committed that crime" and "I feel sad about having committed that crime."

[11] Yet when one pictures guilt prompting tears, it is natural to picture the scene as one where the transgressor is confronted and accused. By contrast, we have no trouble imagining the tears of sadness falling in private.

Consider, second, Prinz's argument that attempts to build moral judgment out of non-moral emotions (for example, blame, which includes "other-directed emotions, such as anger, contempt, disgust, resentment, and indignation" (Prinz 2008a: 368–9)) combined with meta-emotions, third-party concern, and abstract ideas.[12] As a way of testing the adequacy of this empiricist hypothesis, let us imagine someone who satisfies all these components for one of the other-directed emotions that Prinz mentions: disgust. Suppose Ernie sees Bert vomit and feels disgust. Perhaps Ernie feels embarrassed at this response, or perhaps he is pleased with it; in either case, he manifests meta-emotions. When Ernie thinks about some distant other person vomiting, he finds this idea pretty disgusting too; hence the emotion can be directed at third parties. Ernie is also capable of forming abstract ideas, so even the thought of vomit in some abstract sense makes him feel queasy.

It is clear that Ernie is pretty unhappy about Bert's vomiting, but it is considerably less clear that he has made a full-blooded moral judgment about it. We apparently need not credit him with the ideas that vomiting is *wrong*, that Bert has *transgressed*, or that vomiters *deserve* reprimand (or that non-vomiters deserve praise). These, it will be noticed, are distinctly *cognitive* elements that are lacking in Prinz's account. If our conception of moral judgment privileges such cognitive elements, then Prinz's project must be deemed inadequate.

According to Nichols, core moral judgments concern harm prohibitions that are lent resonance and prominence by an innate affective program. One might also want to insist that a key element of moral norms (as opposed to other kinds of non-hypothetical norms) is that they have a special kind of practical authority. Foot, for example, discusses the Kantian idea that to transgress against a moral imperative is irrational, whereas transgressions against etiquette need not be. Elsewhere, I have followed John Mackie (1977) in suggesting that moral imperatives are conceptually "non-institutional" whereas those of etiquette are not (see Joyce 2001, 2011b). Nichols doesn't deny this extra authority with which morality is imbued, but he argues that it comes into the picture later: as a consequence of the affective resonance of this class of norms. He writes that "the affective response seems to play a major role in determining the strength of one's normative commitments . . . [T]he affect-backed norms are treated as having justifications that go beyond the conventional" (Nichols 2004: 159).

But the nature of this connection remains puzzling. It can be granted that emotionally charged norms may be more memorable and seem more important. Yet it does not obviously follow that such resonant norms must also be accorded a stronger binding quality, that they will seem to hold independently of any institutional backing, that they will appear to require no further justification, or that one will be tempted to treat their violation as a form of irrationality. If affectively underwritten norms happen to produce this air of practical authority, then this is a phenomenon requiring

[12] Prinz (2014) adds the capacity for abstraction to his list of general mechanisms that account for moral judgment.

explanation. Until such an explanation is offered, then to the extent that one's conception of a moral judgment makes central this idea of special practical *authority*, Nichols's empiricist hypothesis doesn't pass muster.

The ingredients offered by Sterelny suffice for a social creature who is sensitive to harm situations, who feels empathy for his fellows, who generalizes from exemplars, for whom departures from the cooperative order are memorable and salient, and who, as a consequence, operates extremely well in his social world. But where is the morality? The language Sterelny uses does seem to acknowledge that there is at least some important element of morality that is *more* than the joint exercise of these capacities, for he writes of moral norms "developing from" and being "grafted on top of" these capacities. This seems correct, for it appears no great feat of the imagination to envisage a social creature who enjoys the traits allowed by Sterelny but who is nevertheless constitutionally incapable of making moral judgments concerning an action's meriting punishment, a norm's having convention-transcending practical authority, or even an outcome's being desirable (as opposed to being desired). It is, in other words, not hard to imagine a creature who enjoys all Sterelny's ingredients but for whom full-blooded moral cognition does not simply "develop." Hence, if one's conception of moral judgment privileges such cognitive accomplishments, then what is required is an explanation for why and how it does develop from these ingredients in the normal human case.

The Third Node of Imprecision: Moral Judgment

From this review of some anti-nativist hypotheses, a pattern has emerged. Anti-nativists tend to understand moral judgment in term of emotional traits which, they think, have more general psychological roles and thus are unlikely to count as mechanisms dedicated to the production of moral judgment. However, the ingredients they offer appear to leave certain more cognitive elements of moral judgment unaccounted for. Although I am tempted by the hard-nosed response of insisting that these cognitive components are *essential* to moral judgment and thus that these anti-nativist arguments fail, my considered stance is more pluralistic.[13]

I suggest that the notion of *moral judgment* is sufficiently pliable as to allow different legitimate precisifications. A less demanding conception can be built largely out of emotional resources. To the extent that the less demanding conception might feel unsatisfying, in that it leaves certain cognitive elements of moral judgment unaccounted for, we must recognize the existence of a more demanding conception.[14] It is

[13] In the past I have offered a fairly detailed description of what I take moral judgments to be, involving strong cognitive elements (Joyce 2006a: chapter 2). This characterization has been criticized as being non-mandatory (see Machery and Mallon 2010), and, indeed, Stich finds it necessary to speak of "Joyce-style moral judgments" (Stich 2008: 234).

[14] I should point out that in the interests both of simple expression and playing along with an entrenched dialectic, I am drawing a line between "emotions" and "cognitions" in the orthodox ham-fisted manner. Of

not a matter of there being two or more concepts; it's a matter of there being competing precisifications of the same somewhat indeterminate concept. A liberal conception will count as moral judgments items that the strict conception will not. And even for a paradigm moral judgment about which there is no doubt, the competing conceptions will disagree regarding the criteria in virtue of which the item counts as a moral judgment. It's not a matter of our not knowing which is the correct conception (because we lack data); it's that there is no unique fact of the matter.

A similar view has been expressed in the useful comparison case of the human language faculty. Marc Hauser, Noam Chomsky, and W. Tecumseh Fitch—recognizing that "the word 'language' has highly divergent meanings in different contexts and disciplines" (Hauser, Chomsky and Fitch 2002: 1570)—distinguish between a faculty of language in a broad sense and in a narrow sense. The former, they hypothesize, consists largely if not entirely of capacities that humans share with other animals, whereas the latter (which is basically the capacity for linguistic recursion) is a uniquely human trait.

But whereas Hauser, Chomsky, and Fitch do an admirable job of delineating the various skills and capacities involved in the two senses of "language faculty," I feel somewhat pessimistic that the same can be done for "moral faculty," for here, it seems to me, matters are considerably more nebulous. The three examples of non-nativists described above—Prinz, Nichols, and Sterelny—hardly present a univocal picture of what a liberal conception of moral judgment might look like. They do, very broadly, all think that emotions are terribly important, but beyond this, three noticeably different views are articulated. To the extent that my own views have represented the advocacy of an opposing more cognitivist position, I haven't denied the importance of emotions but have maintained that cognitive components are vital too (cognitive components, that is, for which the anti-nativist proposals do not succeed in accounting). Yet if asked to characterize the crucial cognitive elements of the more demanding conception, I have nothing so simple and distinct as "recursion" to say. Rather, I will point to aspects of moral judgment like *desert, transgression, practical authority* (and so on), and declare (a) that these are *cognitions* (for example, judging that X deserves punishment is not something one just "feels"), and (b) that emotional resources alone do not suffice to account for them. But the answer lacks precision (though is no less reasonable

course, the real distinction is nuanced and complicated. I should also say something to clarify the relation (or lack thereof) between the view under discussion and the literature on the neuroscience of moral judgment, in which the question of emotions versus cognitions looms large. Joshua Greene argues that some moral judgments (deontological ones) stem from emotional arousal, whereas others (consequentialist judgments) flow from rational faculties. (See Greene et al. 2001.) Be that as it may, the deontological judgments that are prompted by emotional responses still, in my book, involve obvious cognitive elements. For example, judging that someone has an inalienable right to something (for which consequentialist considerations are irrelevant) involves the deployment of the hefty abstract concept *inalienable right*. Similarly, Jon Haidt's (2001) work may show that moral judgments are little more than *post hoc* rationalizations of knee-jerk emotional responses, but this should not be confused with the claim that moral judgments are nothing more than emotional responses. Although Greene and Haidt (and others) underline the central role of emotion in moral judgment, they need not be interpreted as proponents of a less demanding conception of moral judgment.

for that): The list of cognitions is worryingly open-ended (note the "and so on"), and, moreover, not one of the items listed is easily defined. The literature on with what kind of "practical authority" our moral norms are invested, for example, stretches back to the ancient Greeks and continues unabated.

Perhaps my pessimism is premature, and distinct senses of "moral judgment" can be delineated with a reasonable amount of specificity. Or perhaps my doubt will be borne out, and the whole concept will remain inchoate and ill defined. In either case, what is evident is that it is a mistake to choose one particular characterization of "moral judgment" and declare it to be the true and unique deserver of that name. I have argued elsewhere (Joyce 2012) that this kind of indeterminacy may span the difference between metaethical cognitivism and noncognitivism, and also the difference between moral realism and moral skepticism. In other words, there may be some legitimate precisification of the concept *moral rightness* (for example) according to which rightness is a real property of certain actions; but there may be other equally legitimate precisifications according to which no such property exists anywhere. How might this sort of indeterminacy affect the debate over moral nativism?

It is possible (and not unlikely) that on *any* precisification of "moral judgment" (and on any disambiguation of "innate") moral nativism is false. But it is also possible that moral nativism is true for certain precifisications and false for others. Certainly the *plausibility* of various pro-nativist and anti-nativist arguments varies according to different conceptions of the target trait. For example, if one is concerned with questions of universality, then the less demanding our conception of a moral judgment, the more likely it is that we will find evidence of universality, since, as a truistic rule of thumb, X + Y is going to occur more often than X + Y + Z. On these grounds, Stich objects that the rich conception of moral judgment that I offered (in Joyce 2006a) spells problems for moral nativism: "For if moral judgment requires *all of that*, what reason is there to think that people in cultures very different from ours *make* moral judgments?" (Stich 2008: 233).[15] If this is correct, then (roughly speaking) richly construed moral judgments are less likely to be universal, thus favoring the non-nativist case (various aforementioned complications with universality aside).

A number of opponents of moral nativism allow that some kind of *normative* nativism might be true. Earlier we saw Nichols accept nativism about non-hypothetical norms. Edouard Machery and Ron Mallon (2010) also accept the plausibility of nativism about normative cognition ("that is, the capacity to grasp norms and to make normative judgments" (2010: 4))—where nativism is understood in evolutionary terms. What they insist upon is that *moral* judgment is but a proper subset of the normative, and there is no evidence for any psychological adaptations dedicated to moral thinking in particular. While it cannot be reasonably denied that the category of the normative is larger than the category of the moral, it should also be noted that *how much larger*

[15] Machery and Mallon make the same point: "[Joyce's] claim is substantive and provocative precisely because of the rich characterization of moral judgments that he offers" (2010: 21).

depends on what conception of the moral one endorses. A demanding conception will make the moral a smaller subset of the normative; a less demanding conception will yield a larger subset. The larger the subset, however, the more plausibility there is to the claim that it is in fact *moral* judgment that is the distinct adaptation, while the human capacity to make non-moral normative judgments is a case of aspects of a biological adaptation being coopted for new uses.[16] (This position will be strengthened if we have a plausible hypothesis about why moral judgment in particular might have been adaptive to our ancestors while lacking a hypothesis about why normative judgments in general might have been adaptive.) Thus, again, a less demanding conception of moral judgment might be more amenable to a nativist explanation than a more demanding one.

On the other hand, POS arguments seem to cut the other way. If a thin moral judgment can be constructed out of evolutionarily preexisting mechanisms, then heaping more demands on the conception of moral judgment ("thickening" it) lowers the probability that these mechanisms will remain sufficient to the explanatory task. Again speaking roughly: Richly construed moral judgments will need more mechanisms to explain them; and the more mechanisms to which one must appeal, the more likely it is that at some point one will need to appeal to a *dedicated* mechanism, thus favoring the nativist case. In this essay I haven't attempted the difficult task of arguing that a POS-style argument is plausible even for a demanding conception of moral judgment (though I admit to some sympathy with the project); my objective is simply to draw attention to the fact that the plausibility of the argument may vary according to how the target trait is drawn.

Conclusion

The upshot is that both moral nativism and moral non-nativism may be perfectly defensible positions, and may remain so even when all data is in. This, I predict, will not be a popular conclusion—philosophers and scientists alike prefer their truths tidier—but it is surely worthwhile to diagnose, in advance, those points of conceptual imprecision that may confound future debate.

[16] I am making a debatable background assumption here: that if trait *T* has adaptive function *Fa*, then, for whatever processes make possible "co-opting" *T* for new functions *Fb, Fc*, and so on, it will be *prima facie* more probable that these processes will have co-opted *T* for fewer new functions than for more new functions. Assessing such a principle would be a complicated task; here I leave it at an intuitive level.

7

Evolution, Truth-Tracking, and Moral Skepticism

Introduction

Genealogical debunking arguments are as old as the hills. In the fourth century BC, Euhemerus sought to undermine religious beliefs by alleging facts about their origins: He suspected that divine and mythological beings were in fact historical persons whose qualities and exploits had become exaggerated and distorted with the passage of generations. Early Christian writers such as Lactantius and Origen deployed euhemeristic arguments discriminately: against paganism and idolatry, but not (of course) against their own religion. Nietzsche, Marx, and Freud (to name but three intellectual heavyweights) all sought to disparage religion by presenting views on its sociological or psychological source.

How exactly are such arguments supposed to work? Nietzsche might have hoped that once Christians had seen their religion labeled "a slave morality"—once they had been persuaded that it was historically rooted in hatred dressed up as love—they would turn away from it blushing. But if such a rejection were to occur simply as a psychological response then it would not be a debunking *argument* at all. Freud seemed to acknowledge this when he wrote that once we recognize religion as a grand wish-fulfillment, "our attitude to the problem of religion will undergo a marked displacement" (Freud [1927] 1987: 215). That may well be so, but, again, if this is no more than a prediction of psychological causal consequences, then one is left wondering whether there is supposed to be an actual *refutation* of theism in Freud's writings. Could one accept all that Freud says about religion as neurotic wish-fulfillment while reasonably maintaining theistic belief?

My interest here is in morality rather than religion. In recent years there has been a burgeoning of interest in genealogical arguments targeting morality, with Darwinian evolution being the genealogy appealed to most conspicuously as responsible for the debunking. The empirical relation between Darwinian evolution and human moral judgment is something about which the jury still deliberates. One hypothesis, called "moral nativism," is that the human mind contains faculties dedicated to moral judgment and forged by natural selection for this task. A competing hypothesis is that moral judgment is a cultural achievement involving an array of psychological faculties that evolved for other purposes. Adjudicating these and alternative views is not the

purpose of this essay. Rather, my goal is to examine whether the truth of the former hypothesis would undermine morality. In one sense, the details of the Darwinian hypothesis are redundant to this task, for, as we shall see, what really matters to a debunking argument is whether our moral judgments are the output of a faculty that, we have reason to suspect, does not track the moral facts. And this might be so even if moral nativism is false; for even if moral judgment is not a biological adaptation, it might nevertheless be produced by psychological faculties that fail to track the truth. For example, Freud's wish-fulfillment theory of religion could be extended quite naturally (and not entirely implausibly) to moral systems, in which case the possibility of a debunking argument would arise without evolution getting a look-in.[1] Nevertheless, I judge that there is good reason to continue to keep the evolutionary nativist hypothesis in clear sight, since (A) there is a decent chance that it is true, and (B) there is a solid case to be made that it implies that moral judgments are the output of a non-truth-tracking process. To this end, I will continue throughout to refer to *evolutionary debunking arguments* (EDAs), though I should like the reader to note that I have resisted a strong temptation to prefer the broader label *genealogical debunking arguments*.[2] One must bear in mind that the evolutionary hypothesis might well be substituted in an EDA by some other kind of genealogical hypothesis.

Types of Debunking

Before examining the logic of an evolutionary debunking argument, let us pause to wonder about the nature of the conclusion. What, exactly, is *debunking*? The answer is that it can be many things, though no doubt the associated arguments vary greatly in plausibility. Here are some potential conclusions to moral EDAs, followed by a quick survey of philosophers who argue for these theses.

1. All moral judgments are false.
2. All moral judgments are false insofar as they involve a claim to objectivity.
3. Certain normative moral theories (such as Kantianism)—but not all such theories—should be rejected.
4. All moral judgments lack justification.
5. All moral judgments lack justification and permanently so.

This is not an exhaustive list, but it suffices for my purposes. Both (1) and (2) undermine certain metaethical theories while vindicating others. The metaethical theory vindicated by (1) is the error theory (see Mackie 1977; Joyce 2001). (2), on the other hand, undermines any metaethical theory according to which there are objectively

[1] In fact, Freud views morality either as the internalization of parental norms in order to resolve the Oedipal complex, or as a means of suppressing aggressive and sexual instincts in order to make possible a cooperative society.

[2] In part this is because "EDA" appears elsewhere in the literature, and I am loath to multiply abbreviations.

true moral claims, which (in many people's books) amounts to saying that (2) undermines moral realism. The claims made in (4) and (5), by contrast, are compatible with the possibility that moral facts exist and objectively so. Therefore any EDA that aims to establish (4) or (5) is not an attack on moral realism.

It is possible that (2) implies (1), but only by assuming a controversial bridging premise: namely, that all moral judgments necessarily include a claim to objectivity. One person who possibly argues in this manner is Michael Ruse (2006, 2009). He speculates that having a faculty that issues moral judgments was adaptive to our ancestors because those judgments strengthened their motivation to cooperate, and that this occurred precisely because those judgments were imbued with objectivity: "The Darwinian argues that morality simply does not work (from a biological perspective) unless we believe that it is objective" (1986: 253). He goes on: "[M]orality is a collective illusion foisted upon us by our genes. Note, however, that the illusion lies not in the morality itself, but in its sense of objectivity" (253). The last comment makes it seem as if Ruse is seeking to establish (2). But elsewhere he appears to endorse the bridging premise that would take him to (1): "Ethics is subjective, but its meaning is objective" (Ruse 2006: 22); "[W]hat I want to suggest is that . . . the *meaning* of morality is that it is objective" (Ruse 2009: 507). If morality is necessarily objective, then establishing that there are no objective moral truths—as (2) claims—entails that there are no moral truths *tout court*—as (1) claims.

Of this strong EDA tentatively attributed to Ruse, two things should be noted. First, it requires a highly contentious bridging premise, which itself cannot be established on genealogical grounds alone. Second, this attempt to establish (1) by EDA would be, to my knowledge, unique in the literature. Contrary to widespread opinion, establishing the error-theoretic thesis that all moral judgments are false is simply not the evolutionary debunker's typical goal. It is frustrating to read a commentator on my own debunking argument interpret the intended conclusion as "*all* moral judgements are false" (Mason 2010: 775).[3]

Sharon Street's EDA seeks to establish (2) while explicitly resisting (1); she aims to debunk not morality but moral realism (Street 2006, 2008). The slide from (2) to (1) is avoided because Street allows the viability of moral *constructivism*, according to which objectivity is not an essential feature of moral concepts. Kevin Brosnan interprets Street as inferring "that our moral beliefs are probably false" (Brosnan 2011: 52), but this is incorrect; the conclusion of her EDA is that *moral realism* is probably false. The

[3] Mason might perhaps be forgiven this infelicity on two grounds. First, I certainly have argued for (1) (see my 2001—where, for trifling reasons, I prefer "untrue" to "false"); I have not, however, tried to establish (1) using an EDA. Second, I did once suggest that the label "error theory" might be expanded so as to denote the disjunction of (1) and (4) (Joyce 2006a: 223); and I then proceeded to attempt to establish (4) by an EDA and called it an "error-theoretic" conclusion. I now think that this suggested expansion of the label was injudicious. The observation made earlier—that (4) is in fact compatible with realism—should suffice to show why. (Thanks to Hatha McDivitt for pointing this out to me.)

clause about "probability" arises because part of Street's EDA is that the moral realist might claim that our evolved moral judgments match the objective moral facts without being causally connected to them, in which case (Street thinks) the realist is committed to "a fluke of luck that's . . . extremely unlikely" (Street 2006: 12).

One finds EDAs with thesis (3) as their conclusion advocated in the work of Joshua Greene (2008) and Peter Singer (2005). Greene and Singer argue that an examination of the evolutionary basis of certain widespread moral intuitions casts their practical relevance into doubt. They go on to argue that certain normative theories—basically Kantian in flavor—draw their appeal largely from their conformity with these intuitions, and thus to undermine this class of intuitions is to undermine this class of theories. Whether they think that this reveals the theories to be unjustified, or false, or probably false, is not entirely clear; hence my preference for the vague phrasing of (3): ". . . should be rejected." Greene allows that Kantian intuitions *might* be correct by "coincidence," but (like Street) he thinks that, given their evolutionary history, it is "*unlikely* that they reflect any sort of deep moral truth" (2008: 71). In any case, neither Greene nor Singer is out to support the error theory; each thinks it likely that many moral judgments remain true, so long as these judgments can be underwritten on consequentialist grounds. It is, furthermore, possible that they can maintain that many moral judgments remain *objectively* true, to the extent that consequentialism allows this conclusion.

Indeed, Guy Kahane (2011) argues that debunkers like Greene and Singer *must* presuppose a kind of objectivism, though I confess to finding his argument rather perplexing. Of EDAs and subjectivist/constructivist metaethical possibilities, Kahane writes that

> if there is no attitude-independent truth for our attitudes to track, how could it make sense to worry whether these attitudes have their distal origins in a truth-tracking process? (2011: 112).

Perhaps I am missing something, but I find Kahane's rhetorical question easy to answer. Consider something which we can all agree has constructivist status: the value of money. A given piece of paper is worth $10 because and only because of some kind of collective decision to treat it as worth $10 (construing the word "decision" very broadly); its value is not an objective matter. Suppose Fred is a newcomer to our country and unsure about the respective values of the various pieces of metal and paper that we use as money; but he is also an idiot, and decides to form his beliefs on the matter on the basis of consulting tea leaves (also known as "tasseography"). Clutching a piece of paper (which happens to be a $10 note), Fred examines the tea leaves and decides that it is worth $10. It is clear that Fred's belief, though true, has something wrong with it. We could, in fact, deploy a kind of genealogical debunking argument against the belief, based on the fact that the belief is the product of a process that doesn't track the truth. And yet we can all be sure (including Fred) that the value of money is not an objective phenomenon. Thus, contra Kahane, subjectivist/constructivist metaethical views may

be just as subject to EDAs as objectivist views. What Kahane seems to overlook is that subjectivist/constructivist facts are *trackable* too, and thus beliefs and intuitions about such facts can be produced by processes that succeed or fail to track them.[4]

What I think one should say about Fred's belief is that it is (though true) *unjustified*. This brings us to thesis (4), which is the conclusion of the EDA that I have myself advocated against morality (Joyce 2006a: chapter 6), and which will be the focus of the following discussion. Contributing to the difficulty of articulating and defending such an argument is the fact that epistemological justification is a disputed concept, and the argument may be more or less plausible according to one's epistemological leanings. The comments of the following brief paragraph, however, I think we can take as common ground.

Many types of things can be correctly spoken of as "justified" or "unjustified"—actions, emotions, plans, laws, and so on—but our current concern is epistemological justification, which applies paradigmatically (and probably solely) to *beliefs*. Epistemological justification is distinct from truth: A belief can be justified but false, or unjustified but true. Epistemological justification is relative: One person's belief that p may be justified while another person's belief that p is not justified. A person's belief that p may be unjustified at time t but later gain justification; or justified at time t but later lose justification. Beyond these and a few other platitudes, a bewildering degree of disagreement reigns.

Justification and Truth-Tracking

A justified belief, if true, cannot be true by accident. What makes it justified is that it connects to the facts in the appropriate manner. (Even this second comment might be rejected by certain epistemological coherentists, but one cannot please everyone, especially if they insist on holding outlandish views.[5]) According to evidentialism, the justifying relation between a belief and the facts is a complicated affair that depends on the possession of evidence in favor of that belief—evidence being something that

[4] Consider, for example, Ronald Milo's contractarian constructivism, according to which moral facts are determined by the choices of a hypothetical idealized group of rational contractors. It is not an objectivist theory, since the instantiation of moral facts depends on some "state of affairs [being made] the object of an intentional psychological state" (Milo 1995: 192). (Milo calls it a "stance dependent" theory.) But it is an "evidence-independent" theory, inasmuch as the moral facts obtain independently of our having reasons or evidence for believing them. Given the latter quality, the moral facts are (according to Milo's theory) things that we can track well or poorly, employing sound or foolish methods.

[5] I made critical comments focused specifically on coherentist moral epistemology in Joyce 2006a: 216–17, which I more or less stand by. A coherentism that aims to forge moral intuitions into a tidy package and declare them thereby justified is one whose flaws are so glaring as to require no further comment. By contrast, a truly *wide* reflective equilibrium, which takes into account empirical evidence and probabilities concerning the origin of those human moral intuitions, has more to recommend it; but the price of plausibility is the possibility of moral debunking. Here I second Singer's thoughts: "If the interpretation is truly wide enough to countenance the rejection of all our ordinary moral beliefs, then I have no objection to it" (Singer 2005: 347).

confers reasonableness or confirmation on a belief. According to the rival reliabilism, the justifying relation between a belief and the facts depends on the existence of a reliable mediating belief-formation process—reliability being a quality that probabilifies the truth of output beliefs. Both families of theories allow that false beliefs may have the characteristic that renders them justified.

In recent years, several critics of EDAs have favored reliabilism as a key part of their critiques (Carruthers and James 2008; Wielenberg 2010); thus, in order to meet these critics on their own ground, for the purposes of this essay I am willing to accept a reliabilist perspective. In the end it makes little difference, since what is really at issue is whether moral beliefs are the product of a truth-tracking process, and recognition of this fact appears to be common ground for both reliabilists and evidentialists.[6] Let us say that a belief-formation process is reliable insofar as it produces beliefs that track the truth. The intuition at the heart of truth-tracking is that beliefs may or may not be *sensitive* to the facts which they represent.

It is important to stress that we are not directly assessing beliefs; the claim is not that a belief is unjustified if it fails to track the truth—for any false belief trivially fails to track the truth, yet nobody wants the conclusion that all false beliefs are unjustified (see White 2010: 580). Rather, the claim is that a belief (whether true or false) is unjustified if it is the product of a belief-formation *process* that sufficiently fails to track the truth (Goldman 1979, 2008). Suppose that Fred forms the belief that *p* on the basis of tasseography, and that *p* happens to be false. The fact that this belief is insensitive to the truth does not render it unjustified; rather, what renders it unjustified is that tasseography is a process that is insensitive to the truth. This conclusion would remain even if the belief that *p* turned out, flukily, to be true.

There is a natural temptation to understand this truth-sensitivity in terms of counterfactual covariation: (i) if *p*, then S believes *p*, and (ii) if *not-p*, then S does not believe *p*. But this interpretation is problematic and is, in any case, optional. One problem for the counterfactual interpretation is to account for beliefs concerning necessary truths or necessary falsehoods, since in these cases the antecedents of (i) or (ii) may count as necessarily false, yielding counterpossible conditionals that are, at first blush, vacuously true. This has been thought to be a particular problem for EDAs concerning morality, for here, it may be argued, beliefs with non-contingent content abound. (See White 2010: 583; Enoch 2010: 433; Wielenberg 2010: 455–6.) The proposition "Promise-breaking is wrong" (perhaps with a "*ceteris paribus*" thrown in) is, some would claim, not merely true but *necessarily* so. This would problematize any attempt to assess the counterfactual "If it were not the case that promise-breaking is wrong, you would still believe that it is." Ruse stumbles into this problem when he writes "Given two worlds, identical except that one has an objective morality and the other does not,

[6] The key difference between the views is that the evidentialist holds that the truth-probabilifying elements of the process must be accessible to the agent. The way I see it, the evidentialist can still maintain that in order to be justified, beliefs must track the truth; it is just that the evidentialist has a particular way of restricting what satisfies this criterion.

the humans therein would think and act in exactly the same ways" (1986: 254; see also Mackie 1977: 49). If some cogent views about moral supervenience hold—that two situations identical in all non-moral properties must be identical in all moral properties—then Ruse's thought experiment is tricky at best.

Moves are afoot to make sense of non-vacuous counterpossibles (Restall 1997; Vander Laan 2004; Brogaard and Salerno 2013), and there does seem to be an intuition in favor of doing so being a desideratum. Compare the following two claims: "If 6 were a prime number we'd have some serious rethinking to do" and "If 6 were a prime number it would make no difference to mathematical calculations." The former seems intuitively to have a claim to truth that the latter lacks. A promising solution is to appeal to impossible worlds, which, while sounding fishy, can be understood in innocuous terms. (I like Ed Mares's description: "If we pay for ideology in the coin of ontology, then the doctrine of impossible worlds can be bought with loose change lying round the house of almost any possible worlds theorist" (Mares 1997: 525).)

However, even if such attempts to vindicate non-vacuous counterpossibles fail, there remains plenty of room for endeavoring to make sense of the idea of processes that succeed or fail to track necessary truths. Consider Fred again, who decides whether or not numbers are prime on the basis of tasseography. Let's say that he forms the belief that 7 is prime. Intuitively, one still wants to say that tasseography fails to track the truth in this case—that it is in some sense *insensitive* to the truth—and one should be able to do so without having to consider propositions like "If 7 were not a prime number, then Fred would still believe it." The crucial characteristic seems to be that mathematical facts about the primeness of numbers play *no explanatory role* in why the tea leaves formed a certain pattern in the cup, which is the sole factor in determining Fred's belief. And this, surely, is something we can understand without needing to assess the truth of counterpossibles.

Here I am agreeing with Gilbert Harman's claim that the counterfactual test of whether *H* explains *E* "is only appropriate in certain contexts" (1986: 63). In the case of moral facts explaining moral judgments, Harman says that rather than an assessment of counterfactuals, "what's needed is some account of *how* the actual wrongness of [something] could help explain [someone's] disapproval of it. And we have to be able to believe in this account. We cannot just make something up" (63). Such an account is what's lacking regarding the relation between numbers being prime and the pattern of tea leaves in a cup.

[Aside: Imagine someone, Mary, being exposed to skewed "evidence" indicating that tasseography *is* an effective way of determining whether numbers are prime. Suppose Mary lived at a time (say, 1,000 years ago) where believing in supernatural forces, and so on, was not unreasonable (in the absence of better hypotheses), where all the epistemic authorities in her community unanimously endorsed tasseography, and so forth. If this story is fleshed out, there comes a point where many (myself included) become inclined to say that Mary's belief that 7 is prime, though formed solely on the basis of tea leaves, is justified. But surely it remains true that 7's being prime plays no

explanatory role in her forming that belief, which seems to conflict with the previous suggestion. The reliabilist might respond that the process of forming beliefs on the basis of tasseography was not really the same process 1,000 years ago as it is now. *Then* someone might consult epistemic authorities, weigh data, examine alternative hypotheses, and so on, and reasonably come to the conclusion that tasseography is dependable. (I am, in any case, supposing this for the sake of argument.) *Now*, by contrast, an ordinarily situated person can give credence to tasseography only by neglecting superior alternative hypotheses and rejecting a great deal of available data about the causal structure of the natural world. If one widens one's attention from the narrow process of *tasseography* toward the broader process of employing techniques that do (or do not) involve consulting one's epistemic community, that do (or do not) involve ignoring available alternative hypotheses, and so forth, then one can say that historical-Mary's belief is the product of a reliable kind of process—whereas contemporary-Fred's belief (with the same content, formed on the basis of the same *narrow* process) is not. Roughly, we can say that Fred "should know better," whereas we cannot say this of Mary. Of course, 7's being prime still plays no role in explaining Mary's belief that 7 is prime, but one might nevertheless maintain that Mary is employing *broad* belief formation practices which probabilify that her beliefs are (A) true, and (B) explicable (in part) by the relevant facts, whereas contemporary Fred is not. Should we, then, construe a person's processes narrowly (according to which Mary's belief is unjustified) or broadly (according to which Mary's belief is justified)? The challenge of answering this question is, I take it, equivalent to the need for the reliabilist to answer the well-known *generality problem*, which I consider a serious obstacle. (See Conee and Feldman 1998.) The ecumenical conclusion that Mary's belief is in one sense epistemically justified and in another sense epistemically unjustified is one from which I don't recoil. (See Goldman 1988.)]

Adaptation and Truth-Tracking

In the context of evolutionary discussion, it is sometimes useful to speak of traits whose function it is to track the truth, which is to say that the trait's meeting a certain threshold of success at matching the facts enhanced reproductive fitness among ancestors (relative to the competition) and thus explains why the trait emerged and persisted. This is intended to apply only to traits that involve *representational* states. Perspiration may covary with bodily temperature, but we don't want to say that perspiration represents bodily temperature or anything else (at least, *I* don't want to say that), and thus the question of whether perspiration "tracks the truth" does not arise. If, by contrast, an evolved psychological faculty produces some species of *judgment*, then the prerequisite of producing representational states is satisfied, and one can then sensibly ask whether the function of the faculty is to track the truth. (I remain uncommitted on the extent of the domain of representational states.) The only "truth" with which we are concerned here is the one that is represented. Suppose an evolved faculty has the

function of producing judgments of the form "X is P." These representational states might covary robustly with (or be explained by) *Y's being Q*, but we would not on that account say that the faculty tracks the truth or has the function of doing so. Whether the faculty tracks the truth depends on whether the judgments covary with (or are explained by) *those fact(s) that they represent*—in this case, X's being P. And whether the faculty *has the function of* tracking that truth depends on whether success at truth-tracking explains the emergence and persistence (and thus the very existence) of the faculty. (When Street talks of a "tracking account" of the evolution of a moral faculty—as opposed to what she calls an "adaptive link account"—she means the hypothesis according to which the faculty *has the evolutionary function* of truth-tracking. See Street 2006 and 2008: 210.)

Some concrete examples might help. Let us suppose that humans have been hardwired by natural selection to perform certain basic arithmetical functions (see Butterworth 1999; Dahaene 1997). Simplifying things for the sake of illustration: Let us suppose that the belief that $2 + 3 = 5$ is an innate adaptation. Having this proposition immediately available at one's mental fingertips, without calculation or hesitation, might well have been useful for our ancestors in a myriad of ways. But such arithmetical beliefs are useful *only if they are true*. By contrast, having the proposition "$2 + 3 = 6$" immediately available at one's mental fingertips is going to lead to all sorts of practical problems. The faculty that produces these innate beliefs, we must conclude, has the function of tracking the truth: It exists in virtue of producing accurate arithmetical representations. (To reiterate the point made earlier: One need not interpret this as requiring that we can make sense of the counterpossible "If $2 + 3 \neq 5$, then . . ."; it suffices that the fact of $2 + 3$ summing to 5 explains (somehow) the faculty's tendency to give the output belief that $2 + 3 = 5$.) Thus the belief that $2 + 3 = 5$—though (we are supposing) the product of a faculty that evolved in order to help our ancestors make more babies—is the output of a reliable process, and thus counts (*ceteris paribus*) as justified.[7]

[7] Justin Clarke-Doane (2012) doubts these claims. His worries about the counterpossible are not unreasonable, but I have already said something about this. (I hope I have said enough to refute his claim that if the counterpossible is unintelligible, "the argument on behalf of Joyce et al. fails" (2012: 329).) Clarke-Doane's worries about the explanatory claim, on the other hand, would, if cogent, do more damage to my argument; but I do not find them cogent. He claims that "for any mathematical hypothesis that we were selected to believe, H, there is a nonmathematical truth corresponding to H that captures the intuitive reason that belief in H was advantageous is plausible" (2012: 332). What he has in mind is that the usefulness of the ancestral belief that $2 + 3 = 5$ (say) might be explained by reference to "(first-order) logical truths regarding objects in our environments (it is conceivable that they would also sometimes be mereological or impure set-theoretic truths regarding such objects)" (332). There is of course a substantive philosophical question of how mathematical facts can explain mathematical beliefs, into which I have no space to delve here. (The only clarification I would add is that this is distinct from the question of how mathematical facts can have *causal* effects.) In reference to the central point being made here, the issue is whether Clarke-Doane's explanation of the adaptiveness of mathematical beliefs by reference to first-order logical truths would render those beliefs *false*. If not, then my claim that such an explanation presupposes that the beliefs are true may be maintained. If so, then I would happily move on to using a more straightforward and less contentious example to illustrate the point that some beliefs are useful only because they are true, regarding which I have little doubt that one could be quickly found.

Compare this with another example. Let us suppose that humans have been hard-wired by natural selection to systematically make unrealistically positive self-evaluations (see Taylor et al. 1988; Alicke 1985; McKenna et al. 1991). People robustly judge themselves better-than-average in all sorts of ways, including supposing themselves to have an above-average ability to resist the temptation to make unrealistic positive self-evaluations (Friedrich 1996; Pronin et al. 2002). Such everyday delusions might enhance physical health or motivate confident participation in social activities. But the beliefs *do not need to be true* in order to accomplish such adaptive ends; indeed, a great many of them *must* be false, since not everyone can be better than average. The faculty that produces these innate beliefs, we must conclude, does not have the function of tracking the truth: It exists not in virtue of producing accurate self-appraisals, but rather in virtue of producing self-appraisals that benefit the agent's physical and/or psychological well-being. Thus the belief "I am a better-than-average driver" (say) is not the product of a reliable process, and thus is unjustified.

The last conclusion is, of course, far too hasty, and requires qualification. A great many people really are better-than-average drivers, and some of them may well have solid ground for believing this of themselves. Perhaps a person has passed various practical tests at driving school, taken courses in driving, and so on, and has repeatedly come top of the class. Such a person seems to have come by her belief via a process that *is* reliable and thus surely has, one might think, a justified belief on the matter. If, however, we take at face value the claim that the belief that she is better-than-average is *hard-wired*, such that she would carry on believing it come what may, then it follows that had the practical driving tests and so forth supplied her with ample evidence that she is in fact a dreadful driver, she *still* would have concluded that she is better-than-average. Bearing this in mind, her belief that she is better-than-average—even when true and accompanied by ample evidence testifying to this fact—does not look so justified after all.

Once this consequence is brought out, it would be reasonable to complain that the idea of beliefs "hard-wired" by natural selection, immune to all evidence or reason, seems artificial and unfamiliar. This may well be so. The point of the above examples is not to be plausible but to illustrate a contrast. A more realistic portrayal of the trait would maintain that while natural selection may have given us a tendency to over-estimate ourselves in various ways—as a kind of knee-jerk default—we are not incapable of forming accurate beliefs about our abilities when presented with conspicuous and incontrovertible evidence.

This point illustrates the important difference between theses (4) and (5) from earlier: the difference between certain beliefs lacking justification and their lacking justification permanently. In the case of inflated self-evaluations, the thesis that seems correct to me is that when the belief is formed as a knee-jerk default, without reflection or proper sensitivity to the available evidence—when, that is, the belief is to be explained largely by reference to the arousal of an evolved *non-truth-tracking* doxastic faculty—then it lacks epistemic justification. But one is not necessarily stuck in that

position. We are creatures with the capacity to bring other psychological faculties to bear on the matter—faculties that can track the truth in a reliable manner—and when these are employed properly, the same belief that was once unjustified may become justified. A person's initial default belief that she is a better-than-average driver may have been without justification, but sensitive consultation of appropriate evidence may provide the belief with its missing justification.

Given this, the strength of this kind of EDA is to establish a challenge, or a burden of proof. The beliefs that have been called into question *may* be rendered justified by the employment of appropriate epistemic mechanisms (they may be "undebunked"), but until that is accomplished they cannot be considered justified. A prominent target of this kind of EDA is, thus, the epistemic conservative, who holds that firmly held beliefs are "innocent until proven guilty." It's not that the proponent of the EDA need maintain that epistemic conservatism is mistaken, but rather declares that evidence that the beliefs in question emanate from a non-truth-tracking source serves as a defeater of any *prima facie* justification which they might have been accorded.[8]

Morality and Truth-Tracking

Here is not the place to present in any detail hypotheses concerning the evolution of the human faculty for moral judgment (see Alexander 1987; Joyce 2006a; Mikhail 2011; Kitcher 2011). It suffices for our purposes to note that none of the examples of such hypotheses mention that this faculty served reproductive fitness via the production of *true* judgments. Most nativist hypotheses suggest that morality plays a vital role in enhancing social cohesion. Perhaps the adaptiveness of moral thinking lies in the fact that judging an uncooperative action to be *forbidden* might engage a more stalwart motivation to refrain—more stalwart even than thinking of that action as against one's own interests. Yet the plausibility of this adaptational hypothesis seems unaffected by whether one is a moral realist or an error theorist—that is, unaffected by whether one thinks that these uncooperative actions *are* forbidden. According to this hypothesis, then, the evolutionary function of the moral faculty is not truth-tracking.

Other moral nativists emphasize the role that moral judgments can play in signaling one's commitment to social projects (Miller 2007; Nesse 2007). Abiding by moral norms frequently involves foregoing some kind of immediate profit, meaning that morality can function as a *costly* signaling device. Costly signals correlate with *honest* signals, since the profits that can be gained by giving a dishonest signal will cease to

[8] There are some substantive options here, depending on what the details are of one's favored epistemological theory. One might say, as I have, (i) that *evidence of* the beliefs' non-truth-tracking source removes the justification that they heretofore had been accorded; or alternatively one might say (ii) that *the fact of* the beliefs' non-truth-tracking source removes the justification that they heretofore had been accorded. Regarding (i), one might say either (ia) that this knowledge renders previously justified beliefs unjustified, or (ib) that this knowledge reveals that the beliefs were never justified in the first place. In the case of (ii) I take it that only the latter option is viable: that the fact of a non-truth-tracking source implies that the beliefs were never justified in the first place.

provide a net gain if the signal is sufficiently expensive to produce (Zahavi 1977; Noë 2001). Thus, if one's reproductive capacities depend on being selected as a partner in various cooperative ventures (hunting, raising a family, and so on), and those doing the selecting will prefer those who are strongly committed to such ventures, then it may be adaptive to advertise one's prosocial allegiance in a costly fashion. Thus, making moral judgments in a sincere manner may be adaptive as a signaling device. Yet, again, one might be convinced of this hypothesis while maintaining an error-theoretic metaethical stance; there is no pressure to assume that the moral judgments need be *true* in order for them to play this adaptive role.

If these hypotheses are correct, then a truth-tracking moral faculty was not selected for. But it does not follow that a truth-tracking moral faculty was not selected *tout court*. Elliott Sober established the useful distinction between a trait's being *selected* and a trait's being *selected for* (Sober 1984). The latter indicates that the trait is the *target* of selection, in that the nature of the trait plays a causal role in the selective process. The former, by contrast, indicates that the trait is a byproduct of the selective process. Example: As whales evolved, the lipid content in their blubber increased; the insulating quality of lipid-rich tissue is what was selected *for*; the exceptional flammable characteristics of the blubber, by contrast, were merely naturally *selected*.

With this distinction in hand, one can see that it remains entirely possible that an evolved moral faculty does in fact track the truth (that this has been selected), even though it does not have this as its evolutionary function (that is, even though it has not been selected *for* this quality). Several critics of EDAs have pressed this point, in different ways, in making their case. Brosnan (2011) suggests the possibility that cooperation with others is morally good. The evolutionary process would explain both why we believe that cooperation with others is morally good (because doing so enhances the tendency to cooperate in an adaptive manner, say) and why cooperation is in fact good (because it tends to promote well-being, say). David Enoch (2010) presents a structurally similar argument. He speculates that survival or reproductive success is morally good, and that Darwinian forces have shaped our moral beliefs such that they often concern actions and events that promote survival and reproductive success. Thus, even if the *truth* of our ancestors' beliefs does not figure in the account of why they were adaptive, nevertheless they *were* (sometimes and non-accidentally) true. Erik Wielenberg (2010) advocates another such argument, conjecturing that natural selection has provided humans with beliefs concerning individuals being surrounded by "a kind of moral barrier that it is . . . illegitimate for others to cross" (444–5). Such a belief might well have been adaptive in various ways. Moreover, the very cognitive capacities that make forming such a belief possible also guarantee (or at least probabilify) that one has such a "moral barrier," thus ensuring the belief's truth. (See also Skarsaune 2011.)

All such strategies can be seen as appealing to a "third factor" to explain the reliable connection between moral beliefs and the facts they represent. These views allow that moral beliefs did not evolve *in order* to track moral truths, but speculate that some

aspect of the nativist genealogy raises the probability that the moral facts (those which the beliefs in question represent) obtain.

One logical possibility is that the evolutionary process that produced the moral faculty was somehow causally influenced by the existence of moral facts. Consider, by analogy, how Euhemerus sought to debunk the state religion. He claimed that on his travels to Crete he had stumbled upon an ancient tomb of a king named "Zeus," indicating that the stories of Zeus-the-divine were no more than exaggerations derived from the exploits of a historical figure. However, putting aside the obvious rejoinder that there might simply be two entities with the same name, one can imagine a religious believer claiming that *of course* Zeus-the-divine would create a kingly tomb bearing his own name, precisely to test the faith of potential doubters. If this hypothesis is given credence, then finding a tomb marked "King Zeus" is no evidence against the existence of Zeus-the-divine—not even if one has direct evidence that the current conception of the god Zeus really does derive from inflated and half-forgotten stories about a real king. The religious believer simply claims that the god Zeus directly or indirectly causally established this evidence. In the same way, one might claim that the existence of moral facts somehow causally influenced the evolutionary genealogy. I mention this just for completeness; it is hard to imagine anything plausible being made of the idea. (See White 2010: 582–3.)

A more promising possibility is that the moral facts might be identical to, or supervene upon, the very facts described in the genealogical account. This is the strategy that all the aforementioned opponents of the moral EDA adopt. For example, the nativist hypothesis more or less explicitly mentions types of action that conduce to cooperation, and the anti-debunker can claim that whatever conduces to cooperation just *is* morally good. The evolutionary hypothesis might not explicitly invoke moral facts, but the moral facts might be implicitly presupposed by the hypothesis. In the same way, one might explain an alpine avalanche using terms like "snow" and "sunshine," while not mentioning hydrogen atoms at all, but it is not as if this demonstrates that hydrogen atoms played no role in the avalanche; an explanation couched in terms of "snow" implicitly concerns the activity of hydrogen atoms. The vernacular explanation might be more suitable to a given conversational context than the molecular explanation—or vice versa—but neither is superior in the sense of excluding the truth of the other.

On an earlier occasion I supplemented my version of the EDA with general metaethical arguments against moral naturalism, in order to cast doubt on this latter possibility (Joyce 2006a: chapter 6). I do not pretend that those arguments were comprehensive or presented a refutation, and in any case here is not the place to repeat them. Rather, I would like to draw attention to the logic of the dialectic, lest it be misunderstood. As I have stressed, the EDA that I favor represents a *challenge*; it says that the fan of morality has some work to do if justification is to be established or reinstated. Gestures toward a "third factor" do not demonstrate some gaping hole in the EDA that has previously gone unnoticed, but rather represent attempts to meet the challenge that the EDA poses. But this also explains why I find these attempts so unsatisfying, for

none of them seriously undertakes to argue for the connection between moral facts and the evolutionary process in any detail; they rather gesture—sometimes astonishingly vaguely—at a property that *might* suffice to establish the connection, as if showing the mere possibility should demolish the EDA. But the advocate of this kind of EDA was never under the illusion that establishing such a connection is *impossible*; the possibility that moral facts might find a place in the evolutionary genealogy was always acknowledged. What it comes down to is whether this mere possibility can be made *plausible*, and as far as I can see, not one of these debunkers of debunking has made a serious effort.

A slightly cruel comparison might be with those nineteenth-century Christian apologists who tried to explain away the growing evidence of the ancient age of the Earth which seemed to contradict the Biblical texts.[9] Someone who sensibly takes the fossil record as evidence of the venerable age of the Earth need not deny the *possibility* that it was all created by God 6,000 years ago (with fossils placed here to test our faith). Nor need such a person be disturbed by the coherent articulation of a hypothesis according to which dinosaur fossils and a young Earth are co-possible. All that really matters is whether that hypothesis has any plausibility.

By the same token, I am not much disturbed when someone says that *if* the moral facts were thus-and-so, then our ancestors' moral beliefs, though an adaptation, would have reliably tracked those facts. My response is: "OK, show me the theory; let us subject it to critical scrutiny." Wielenberg thinks he can help himself to a certain view of moral facts without pausing to make the view plausible, since he reads me as claiming that even if moral facts existed, *and whatever their nature*, the EDA would reveal our beliefs about them to be unjustified. Although I don't accuse Wielenberg of begging the question, I find his interpretation of the rules of engagement rather unfair. I certainly don't think that the EDA works against *any* construal of moral facts—even fantastic and gerrymandered ones. One might easily stipulate that moral goodness is whatever conduces to genetic replication, and then any evolutionary explanation of anything would of course implicitly involve moral facts. The onus, as I have emphasized, is on making this stipulation plausible. The role of the EDA is to place the burden of proof onto the shoulders of those who believe in justified moral belief. Perhaps Wielenberg's theory can be made to fly; perhaps it cannot. (Hint: I have a raft of objections, should the argument ever get that far.) If he and other opponents of the EDA see themselves as attempting to discharge this burden, then well and good—the EDA will have done its job and the discussion can continue. But if they see themselves as showing that the EDA never posed a challenge in the first place, then I protest.

Brosnan points out that "third factor" responses to the EDA challenge don't actually need to support a kind of naturalism according to which moral facts are identical to or supervene upon natural facts; all that is really needed is that the obtaining of the

[9] For example, Philip Gosse's attempt to explain away all those mysterious Bible-debunking fossils, in his *Omphalos* (1857).

natural facts cited in the genealogy raises the probability of the obtaining of the moral facts (2011: 61). This is correct; but I disagree with Brosnan's diagnosis that this broadening of the requirement reveals that the challenge poses only "a very modest requirement" (61). The broader requirement of *probabilification* may be logically weaker than the requirement of *identity or supervenience*, but it is very far from obvious to me that it is any easier to satisfy.

Something more needs to be said about how the challenge is expressed, for this depends on to whom one takes oneself to be speaking. Against the epistemic evidentialist, the EDA says (roughly) that until a plausible justifying account is presented, moral judgments must be considered unjustified. Against the epistemic reliabilist, by contrast, the EDA says (roughly) that until plausible evidence of a reliable process is presented, we *don't know* whether moral judgments are justified. On previous occasions I pictured my opponent principally as the evidentialist, and thus expressed the EDA in the former manner. The burden for the evidentialist is to provide a theory that *instates* or *reinstates* justification. But the reliabilist raises the possibility that our moral judgments might be the product of a reliable process (whether we know it or not) and thus that our moral judgments might have been justified all along and remain so. The burden for the reliabilist is to provide a persuasive account that establishes that moral judgments are indeed the product of such a reliable process.

This may seem like something of a climb-down for the proponent of the EDA, for he or she appears to have moved from asserting that moral judgments *are* unjustified to asserting that *we don't know* whether they are justified. But it's really not a substantive weakening at all. The skeptical position doesn't change; it's just that how it is expressed needs to be tailored to one's interlocutor. In other words, it might seem that to (1)–(5) from earlier we should add another possible debunking conclusion:

6. We don't know whether moral judgments lack justification,

but in fact this is better thought of as (4*): that is, as (4) articulated for the reliabilist audience. That the proponent of the EDA hasn't really backed down from anything is evidenced by the fact that as soon as his or her interlocutor is again an evidentialist, then he or she will revert to asserting (4).

Conclusion: Shifting the Burden of Proof

Some philosophers will maintain that they can provide positive considerations indicating that many moral judgments connect to the facts in a manner that renders them justified. But even they cannot ignore the EDA, inasmuch as it represents the presentation of a new hypothesis about the place of moral judgments in the world (one, moreover, potentially with empirical backing), and thus any advocate of a truth-tracking hypothesis must either establish the superiority of his/her view over the new hypothesis or demonstrate that the two hypotheses can be jointly accommodated. Either way, any such advocate has some explaining to do.

Numerous other philosophers—and, I hazard to suggest, most of the folk—do not attempt to provide positive justifying considerations for morality, but nevertheless proclaim justification on the basis of some more or less clearly delineated principle of conservatism. It is to this position that the EDA does the most damage. The epistemic conservatist claims that no positive supporting considerations need be articulated in order for moral beliefs to be justified; rather, they receive justification in virtue of (A) the mere fact that we have these beliefs, and (B) the absence of defeating considerations. To my thinking, the provision of evidence that the moral faculty is the product of a process that the moral error theorist can happily endorse represents a significant defeating consideration. It is a *prima facie* defeater, since a persuasive articulation of one of the aforementioned strategies (for example, that the moral facts are identical to certain natural facts) would defeat the defeater.

Conservatism in moral epistemology is rampant. Some explicitly embrace it (such as Huemer 2005; Lycan 1986), but even among those who eschew the label, conservatism is often quietly at work in the background. Roger White tries to nullify EDAs by expressing doubt that people typically take the fact that they believe something as crucial evidence for its truth (2010: 585), but while I accept that people do not typically *explicitly* do so, I am very far from convinced that this phenomenon is not extensive. A person will ordinarily take the fact that she believes there to be a cat in front of her as fairly important evidence that there is a cat. This is revealed by the fact that if the reliability of the belief is called into question—if, say, evidence comes forth that her belief is the product of hypnosis—then a rational person will become considerably less confident that there is indeed a cat in front of her. It is not unreasonable for us to give some initial credence to our beliefs, especially if the practice of doing so has generally withstood the trials of everyday life, but the revelation of an unexpected alternative explanation for a token or type of belief alters the epistemic landscape. A person may then seek to reinstate confidence in her belief that there is a cat before her—by bringing forth other evidence, such as fur-balls and dead mice. But if it is revealed that the beliefs concerning all the "cat evidence" are also the product of hypnosis—such that not only does the belief "Here is a cat" have an alternative explanation, but so too does the belief "Here is a fur-ball," and so forth—then reinstating confidence in the belief may be a trickier affair. Perhaps one might instead investigate the nature of the process of hypnosis—hoping to discover, say, that the hypnotist was the benign sort who instills in his subjects only *true* beliefs—but if no data is forthcoming on that count, then it will be of no avail.

This, I believe, is what nativism is likely to reveal about our moral judgments. First, there is no evidence that can be gathered concerning the obtaining of moral facts that is not itself mediated by the very moral judgments that are in question. We cannot "triangulate" on moral facts using faculties independently of our moral faculty; the very urge to try to do so is driven by intuitions that are themselves the output of the moral faculty. Second, there seems no ground for optimism that the evolutionary process that produced the moral faculty will turn out to be the benign sort that happens to track truths

(even if it lacks this as its adaptive function). Of course, it *might* be—just as the hypnotist *might* have been benign—but epistemic justification requires more than a vague "might."

The first observation is one that can be made with no particular reference to evolutionary matters. It is no great news that the epistemic status of moral judgments is suspect even before Darwinian considerations enter the conversation. But revelation of a non-truth-tracking function for the moral faculty contributes significantly to the second observation, because the expansion of the space of hypotheses forces those inclined to trust their moral beliefs to articulate a persuasive defense of their preferred hypothesis in the face of new competition. If our beliefs in general have tended to steer us right as we navigate the world, and if one considers moral beliefs as in the same boat, then it would not be unreasonable to have some optimism regarding the probable truth of moral beliefs. Moral nativism requires us to consider moral beliefs independently—as not in the same boat—as the output of a psychological faculty with a distinctive evolutionary trajectory. It provides us, moreover, with an explanation of how such a faculty might come to exist while generating systematically false beliefs; and, furthermore, an explanation for why such a body of false beliefs might seem so compelling and almost beyond question. It has not infrequently been claimed against the moral skeptic that one's confidence in fundamental moral intuitions must be far more robust than one's confidence in any obscure philosophical argument (see, for example, Huemer 2005: 116–17). An EDA has the strength to overturn this comparison, by presenting evidence to account for those fundamental moral intuitions—which can account even for their persuasive felt quality—that is compatible with their falsehood (in the sense that even an error theorist can accept the evidence).

I started out this essay mentioning Nietzsche's, Marx's, and Freud's debunking efforts, pondering whether they should be taken as offering a *refutation* of theism. Could one accept all that Freud says about religion as neurotic wish-fulfillment, I asked, while reasonably maintaining theistic belief? And the answer is: "Not easily." Freud's analysis of the psychology of religion would (if plausible) enlarge the space of explanatory hypotheses, banishing the legitimacy of any kind of complacent optimism regarding theistic belief. The analysis would (if plausible) require the believer to earn the right to his or her belief, to articulate positive grounds for belief in the face of an alternative hypothesis (or show that the hypotheses are not true alternatives at all). A Darwinian analysis of moral belief should have the same effect of arousing suspicion about morality, to the extent that the burden lies on the shoulders of the anti-skeptic to articulate a persuasive defense.

8

Irrealism and the Genealogy of Morals

Introduction to Moral Debunking Arguments

A genealogical debunking argument of morality takes data about the origin of moral thinking and uses it to undermine morality. The genealogy could be ontogenetic (like Freud's) or socio-historical (like Nietzsche's or Marx's), but the focus of recent attention has been the evolutionary perspective. "Debunking" and "undermining" are intentionally broad terms, designed to accommodate a number of different strategies and conclusions. Sharon Street's debunking argument, for example, aims to overthrow moral realism, while leaving intact the possibility of non-objective moral facts (for example, those recognized by a constructivist) (Street 2006). Michael Ruse's earlier debunking argument often looks like it has the same aim as Street's, though on occasions he appears to try for a stronger conclusion: that all moral judgments are false (Ruse 1986, 2006, 2009). My own debunking argument has an epistemological conclusion: that all moral judgments are unjustified (Joyce 2006a; Essay 7, this volume).

Calling all of these conclusions instances of "debunking" is, in some sense, prejudicial. The rejection of moral realism, for example, counts as a *debunking* of morality only if one thinks that realism is somehow the natural interpretation of morality—and that is far from obvious. Any act of debunking is at the same time a *vindication* of something. For example, to show that all moral judgments are false would be to vindicate the error-theoretic metaethical view. But I will let this pass, and allow "debunking" to remain as a usefully vague intuitive term for these arguments.

What these disparate arguments often share is a presupposition of cognitivism. Moral judgments can be all false only if moral judgments are the kind of thing that can have truth value. Moral judgments can be all unjustified (in an epistemic sense) only if moral judgments are beliefs. In other words, the noncognitivist—who holds that moral judgments (as mental states) are not beliefs and (as speech acts) are not assertions—will survey the debate over these debunking arguments with an unperturbed air.[1]

[1] While Street's argument does not presuppose cognitivism, at the same time noncognitivists needn't be worried by the argument, since for them the refutation of moral realism will be entirely welcome.

As already mentioned, the debunking argument that I have advocated (and thus, obviously, think most promising) is one with an epistemological conclusion. It is not my intention to defend or elaborate this argument further on this occasion, though it is probably best if I rehearse it briefly in order to provide a comparison class. Rather, in this essay I want to explore two different debunking avenues. First, I shall investigate what the prospects are for a debunking argument that aims to establish a moral error theory. Second, I shall question whether the noncognitivist is warranted in his/her complacency; perhaps a debunking argument against noncognitivism could be developed. My objectives are diagnostic rather than promotional, thus my conclusions regarding both these avenues will be non-committal. Given this, beginning with a brief look at a kind of debunking argument that I think likely to succeed will provide a useful backdrop.

Epistemological Debunking

Recent years have seen a burgeoning of discussion about the evolutionary origins of the human moral faculty.[2] Part of any such nativist explanation must be an account of what it was about moral thinking that served the reproductive purposes of our ancestors. On this point hypotheses diverge, but on most accounts moral thinking was advantageous because it in some manner enhanced their cooperative tendencies. What is striking about these nativist hypotheses is that they seem entirely compatible with the error-theoretic stance; they do not appear to imply or presuppose that any of our ancestors' moral judgments were *true*.

This is not so of evolutionary explanations of any kind of judgment. For example, humans quite possibly have an adaptive mechanism for distinguishing faces from other visual stimuli. But if one were to be (bizarrely!) an error theorist about faces, then the evolutionary explanation for why it might have been useful for our ancestors to have this mechanism would surely fizzle. By contrast, the evolutionary hypothesis that moral thinking emerged because it strengthened social cohesion is no less plausible for the error theorist than anyone else. The best explanation of the face-identifying adaptation classifies it as a *truth-tracking* mechanism; the best explanation of the moral faculty does not classify it as a truth-tracking mechanism. This, it would seem, has epistemological consequences.

Most epistemological theories (and, I am tempted to add dogmatically, all sensible epistemological theories) hold that a belief's being justified depends on its standing in one or other specific relationship to the fact that it represents. To discover that a belief does not stand in this relation to the relevant fact is to discover that the belief lacks justification. (Whether it shows that the belief has lost its justification, or shows that it was never justified in the first place, depends on which family of epistemological

[2] See Alexander 1987; Irons 1996; Krebs 2005; Nichols 2005; Dwyer 2006; Machery and Mallon 2010; Mikhail 2011; Kitcher 2011.

theories one favors.) If the evidence were to come down in support of moral nativism, then this would seem to be confirmation that our moral beliefs have their origins in a process that is not designed for truth-tracking.[3] Note that this would not be a matter of conjuring up a far-fetched unfalsifiable skeptical hypothesis according to which our moral beliefs are bogus (like Descartes' demon); it would be the confirmation of an empirical hypothesis that appears compatible with the systematic falsehood of moral judgments. Such a confirmation, I claim, undermines the epistemic standing of moral judgments.

Justification, of course, is a relative affair. My belief that *p* may be justified while your belief that *p* is not. Perhaps at an earlier time my belief that *p* was also unjustified; perhaps in the future it will become unjustified again (if, say, I ignore mounting evidence against the belief). Thus the conclusion that all moral beliefs are unjustified should not be interpreted as making a stronger claim than is reasonable. The proposition that a belief is unjustified does not exclude the possibility that justification can be attained or reinstated in the future. The force of the epistemological debunking argument is to issue a challenge, to shift a burden of proof.

It is often claimed that the fact that skepticism (about any object of everyday belief) cannot be refuted does not thrust that skeptical stance upon believers, so long as the non-skeptical position also cannot be refuted. Thus it is claimed that the skeptic shoulders a burden of proof: It is not enough to make skepticism irrefutable, the skeptic needs positive arguments against belief. In the event that neither the skeptical nor non-skeptical position is refutable, the non-skeptic can happily carry on with his or her everyday beliefs.

Moral nativism promises to upset this picture by providing a new hypothesis about the place of moral judgments in the world (one, moreover, potentially with empirical backing). Even those who were confident that their moral beliefs are true cannot ignore the evolutionary debunking argument, inasmuch as it is incumbent upon them either to establish that the nativist hypothesis is false or to demonstrate that moral beliefs are true even according to that hypothesis. Either way, they have some work to do. To maintain confidence in moral beliefs in advance of this work is epistemically negligent; any principle that allows one to do so is gullibility dressed up as a methodology.

When I presented this argument on an earlier occasion, I made the rash decision to label it an error-theoretic conclusion (Joyce 2006a: 223). I did this via suggesting that the label "error theory" might denote a disjunction of metaethical positions: either the view that all moral judgments are false or the view that all moral judgments are unjustified. I now recant this suggestion for the following reason. Suppose

[3] Note that "truth-tracking" can be understood epistemically or evolutionarily. The latter refers to what a psychological faculty is *supposed to do* (in evolutionary terms). The former is often taken to refer to a covariation between a belief and the fact that it represents. In fact, I think epistemic truth tracking is quite difficult to spell out, and the covariation analysis runs into difficulties when beliefs concern necessary truths and necessary falsehoods. See Essay 7, this volume, for discussion.

all moral judgments are unjustified. This is consistent with moral judgments being true, and, moreover, objectively true; thus the claim that all moral judgments are unjustified is compatible with moral realism. But the error theory had better not be compatible with moral realism, therefore the view that all moral judgments are unjustified had better not be sufficient for an error theory.[4] It is preferable to keep our metaethical theories separate and be clear that the conclusion to this debunking argument is epistemological in nature. The thesis that all moral judgments are unjustified lacks a label, though it is perfectly acceptable to call it a version of moral skepticism.[5]

I think the epistemological debunking argument outlined in this section has legs. But the benefits of establishing the error theory by stretching the extension of the label in the manner just described (and just renounced) are, to quote Russell, the advantages of theft. I turn now to exploring the prospects of using a debunking argument to establish the moral error theory through honest toil.

Error-Theoretic Debunking

Certainly there are circumstances where learning about the origin of a belief can reveal that belief to be false. My belief that hypnosis cannot instill genuine beliefs in people is falsified if I discover that I was caused to have this belief through hypnosis. But clearly nothing so swift and sneaky as this is going to work in the case of moral judgments and moral nativism. The moral judgment that promise-breaking is wrong, say, simply doesn't imply anything about its own origins in the way that the belief about the limits of hypnosis does. Rather, we shall see, the error-theoretic debunking argument depends on a principle of parsimony.

Let us start with Street's debunking argument, whose conclusion is that moral realism is probably false. She argues that the moral realist, confronted with the truth of moral nativism (we are imagining), faces a dilemma concerning the relation between our moral judgments (products of the distortions and contingencies of our evolutionary ancestry) and the supposed realm of objective moral facts. On the one hand, if there is no relation then it would be an astonishing coincidence if many of our moral judgments were even approximately true—a conclusion supposedly disagreeable to the realist. The problem with the other horn of the dilemma is that it is, according to Street, empirically dubious. I have already noted that the usual nativist hypotheses see the ancestral adaptive pay-off of having a moral faculty in terms of enhancing certain cooperative tendencies, not in terms of tracking moral truths. Street thinks this

[4] Analogy: Ancient Greek atomists didn't have any real evidence in favor of their view; it is not unreasonable to claim that they lacked justification for their beliefs. Yet it would seem weird to be an error theorist about their atom discourse. After all, broadly speaking they got it right!

[5] Academic skepticism about morality is the denial that moral knowledge exists. If knowledge requires both truth and justification, then the error theorist's denial of truth counts as skepticism, as does the epistemological denial of justification. (See Sinnott-Armstrong 2006.)

"adaptive link hypothesis" is superior to any truth-tracking hypothesis for three reasons: It is more parsimonious, it is clearer, and it is more illuminating of the phenomenon it seeks to explain (2006: 129). Street's irrealist conclusion might be put as follows: "There are no objective moral facts." Yet she doesn't deny the possibility of moral facts—they will simply be of a constructivist nature.

What good, one might ask, is this to an error theorist? Let me approach this by quickly comparing Ruse's argument. Ruse maintains that being imbued with a kind of objectivity is the whole point of moral thinking, evolutionarily speaking. Morality serves its adaptive function of strengthening our motivation to cooperate by seeming to be imbued with a kind of inescapable external prescriptivity. "It is precisely because we think that morality is more than mere subjective desires that we are led to obey it" (Ruse 1986: 103). But, Ruse argues, this objectivity is an adaptive illusion. He argues for this latter claim via an implicit appeal to parsimony: Once we have explained why morality *seems* to be objective, there is simply no call for any further explaining in terms of positing a realm of objective moral facts. At this point the conclusion to Ruse's argument looks very similar to that of Street's, reached by somewhat different means. He writes: "[M]orality is a collective illusion foisted upon us by our genes. Note, however, that the illusion lies not in the morality itself, but in its sense of objectivity" (1986: 253).

However, Ruse's discussion contains elements that aren't present in Street's thinking, opening the door to the stronger error-theoretic conclusion. For a start, his emphasis on the adaptive importance of the *objectivity* with which moral prescriptions are infused is not something Street mentions. A strong thread running through his argument is that moral realism is written into the phenomenology of moral experience. But he goes further, apparently moving from phenomenology to semantics: "Ethics is subjective, but its meaning is objective" (Ruse 2006: 22); "[W]hat I want to suggest is that . . . the *meaning* of morality is that it is objective" (Ruse 2009: 507). The move from phenomenology to semantics is not something to which one can help oneself for free, but at the same time it's not unreasonable to assume that the *meaning* of a term is going to reflect our experience of the phenomena denoted by that term. If humans are designed by natural selection to experience morality as objective, then this perhaps makes more plausible the already not-ridiculous thesis that objectivity is an essential quality of morality, conceptually speaking. With this thesis operating as a bridging premise, one can get from the sub-conclusion "There are no objective moral facts" to the conclusion "There are no moral facts." (The two propositions would stand in the same relation as "There are no four-sided squares in the box" and "There are no squares in the box" stand in.)

This bridging premise is a key part of this error-theoretic debunking argument. Street rejects it, hence her conclusion is not error-theoretic. And of course it is an extremely controversial thesis, over which much metaethical ink has been spilt. Part of the problem is that the term "objectivity" is not well defined, and it gets used differently in different areas of philosophy. (For discussion see Joyce 2007a, 2009a.) The notion

that Ruse seems to have in mind is that of moral prescriptions having a kind of external authority: We feel bound to follow them because we experience them as not of our own making (unlike, say, the non-objective prescriptions of fashion).[6] Many philosophers will agree with Ruse that we tend to experience moral norms in this manner, though only some of them (a good number, to be sure) will go along with the stronger claim that this kind of objectivity is *essential* to morality, such that a normative framework stripped of this objectivity wouldn't even count as a "moral" system. Those that do support the stronger semantic claim include both realists (who think that this objectivity can be satisfied) and irrealists (who think that it cannot be satisfied).

Ignoring, for a moment, the difficulty of establishing this bridging premise, let me try to reconstruct the argument that employs it. Whether this actually reflects Ruse's reasoning is not my primary concern, but I will continue to attribute it to him if only for the sake of argument. The argument turns on the application of a parsimony principle:

1. Objective moral facts aren't required to explain anything.
2. If some type of fact plays no explanatory role, then this is ground for disbelieving in this type of fact.

There are deep questions to be raised about both these premises, which I shall turn to in a moment, but initially I want to discuss them just sufficiently to motivate the need for a third premise.

In a sense, nothing is required to explain anything. What I mean by this quizzical claim is that one always has choices in how to explain any phenomenon. If the cat knocks over the vase, one can always explain the broken vase without employing the concept *cat*. Instead of using biological or zoological categories, one could (in principle) make reference to a conglomeration of organic chemicals moving about the room, or a swarm of particles and energy. Thus the concept *cat* isn't required in any explanation of anything. But this hardly means that cats are explanatorily impotent. The crucial point is that cats are reducible to entities that are described at other theoretical levels: chemistry or physics, for example. Thus, even if it were true that reference to objective moral facts isn't needed to explain anything, it wouldn't follow that objective moral facts are explanatorily impotent. For this conclusion a further premise must be added:

3. Objective moral facts are not reducible to any facts that do have explanatory roles.

These premises yield the sub-conclusion:

4. Therefore, there is ground for disbelieving in objective moral facts (that is, there is ground for rejecting moral realism).

[6] This appears to be how Maurice Mandelbaum (1955: 50) uses the term "objectivity."

We can now add the bridging premise:

5. Morality is essentially objective.

And the error-theoretic conclusion follows:

6. Therefore, there is ground for disbelieving in moral facts.

Every single one of the premises is problematic. Let us start by considering premises 1 and 3 together. Ruse's argument for premise 1 is often presented via an analogy (Ruse 1986: 256–7, 2006: 22–3, 2009: 504–5). He refers to the spike of interest in séances in Europe in the aftermath of the First World War. Imagine a grief-stricken mother attending such a séance, during which time she comes to believe that her dead son has spoken to her from beyond the grave. We can explain everything that needs explaining about this belief by reference to psychological and sociological factors; there is no need to suppose that the belief might be *true*. Similarly (Ruse thinks), moral nativism explains everything that needs explaining about why humans judge certain actions to have objective moral status; there is no need to suppose that these judgments might be *true*.

The weakness of the analogy is brought out when we attend to premise 3. In order to suppose that the mother's belief is true, we would have to presume that the world contains supernatural forces, post mortem consciousness, ghosts, and so on—that is, some pretty spooky ontology. It is far from obvious that this is what is required to suppose that judgments about objective morality are true. Moral naturalists (of an objectivist stripe) will often identify moral properties with naturalistic properties that we already accept in our ontological scheme. A utilitarian, for example, may identify moral goodness with happiness.[7] By contrast, any attempt to identify, say, *ghosts* with some cluster of naturalistic properties looks hopeless. In other words, the analog of premise 3 for ghosts looks obviously true. But premise 3 as it stands for objective moral properties will be doubted by many, and therefore cannot stand without argumentative support.

Rather than return attention to the bridging premise 5, let us consider dropping all mention of objectivity, which would allow premises 5 and 6 to evaporate. The revised argument is as follows:

1*. Moral facts aren't required to explain anything.
2. If some type of fact plays no explanatory role, then this is ground for disbelieving in this type of fact.
3*. Moral facts aren't reducible to any facts that do have explanatory roles.
4*. Therefore, there is ground for disbelieving in moral facts.

[7] One may wonder what is objective about something so obviously mind-dependent as *happiness*. But this misidentifies the point. The question is whether the relational proposition "Goodness = happiness" is true objectively (like "Water = H_2O") or true in virtue of some human decision. (See Shafer-Landau 2007: 157–8.)

The stripped-down argument seems a lot like one that Gilbert Harman famously uses to frame his discussion (1977). Harman doesn't endorse the argument, though; he rejects premise 3*, arguing that moral facts are reducible to facts about what reasons we have for acting, which (he thinks), properly understood, are empirical phenomena. Nor does Harman place any emphasis on moral nativism, which for Ruse is the main consideration lying behind the first premise. Harman, rather, appeals to developmental factors to explain how moral judgments might arise from non-truth-tracking mechanisms. This difference doesn't matter to our current concerns; what is significant is that moral judgments can be genealogically explained in a way that makes no reference to their being true. This supports the first premise presumably in the following manner. If moral judgments can be fully explained without reference to moral facts, then this casts immediate doubt on whether moral facts are needed to explain *anything*. (Likewise for Ruse, *mutatis mutandis*, concerning *objective* moral facts.) It seems to me that this move is reasonable, for what possible instance would we recognize of a moral fact playing a role in explaining phenomenon X, where this act of recognition did not involve the use of a moral judgment? Moral facts appear to have what Crispin Wright calls "narrow cosmological role" (1992): Their causal impact always involves someone's having made a judgment concerning their presence. (Cats, by contrast, have wide cosmological role, affecting the world in a myriad of judgment-independent ways: meowing, casting shadows, producing kittens, knocking over vases.) If moral explanations (such as "Fred broke the promise because he's wicked") always depend on someone's having made a moral judgment, but moral judgments can always be fully explained without reference to moral facts, then the explanatory potency of moral explanantia (such as Fred's wickedness) is an illusion.

Whether moral facts can be reduced to facts that do have an explanatory role—as 3* denies (but Harman affirms)—is a matter I don't have space to address here. Ruse (so far as I know) doesn't explicitly argue in favor of premise 3, but I have already noted that its lack of support makes the argument that I'm attributing to him vulnerable. The general format of a defense of premise 3* would be to identify some indispensable feature of moral facts that no naturalistic facts can satisfy. (I am here assuming that facts with explanatory roles must be naturalistic facts.) There are a number of promising contenders for this "indispensable feature," the obvious one being something to do with the categorical practical authority (the "must-be-doneness") of moral facts. Harman reduces moral facts to facts about reasons, and thereby, arguably, satisfies a desideratum of *practical authority*—for what could have more practical authority for a person than her *reasons* for acting? On the other hand, however, Harman thinks that the only viable account of reasons is one that renders them relativistic. Yet one may argue that some quality of absolutism is an "indispensable feature" of moral facts, and if this is correct then Harman's attempt to overthrow premise 3* must be rejected.

I don't propose to spend more time assessing the third premise, for it is premise 2 that should really be occupying our attention in evaluating the error-theoretic debunking argument.

Harman's presentation of the argument does not explicitly endorse premise 2. Summing up his argument (before embarking on his rejection of 3*) he writes that "it remains problematic whether we have any reason to suppose that there are any moral facts" (1977: 23). Imagine it turns out that we do not have any reason to suppose that there are any moral facts. This wouldn't automatically amount to our having a reason to suppose that there are *not* any moral facts. The crucial difference is between premise 2 and the weaker 2B:

2. If some type of fact plays no explanatory role, then this is ground for disbelieving in this type of fact.
2B. If some type of fact plays no explanatory role, then we have no ground for believing in this type of fact.

Premise 2B is more plausible than 2, but it is premise 2 that's required to secure the error-theoretic conclusion. Premise 2B, by comparison, looks like it will feed into an epistemological debunking argument. One cannot derive 2 from 2B without violating the adage "Absence of evidence is not evidence of absence."

But the adage is not to be taken as gospel, for there are certainly circumstances where absence of evidence *is* evidence of absence: most obviously, conditions in which one could reasonably expect to have evidence (see Sober 2009a: 64). For example, if there were a leopard hiding in this room somewhere, it would be reasonable for me to expect to encounter some evidence of the fact; the absence of any such evidence provides evidence of a leopard's absence.

The key question, then, is whether these kinds of circumstances obtain for the case of moral facts. Is it reasonable for one to expect that if there were moral facts we would have evidence of them? I find this a very difficult question to answer, and I suspect that different philosophers will give different reactions. There does seem to be something unsettling about the idea of a realm of moral facts for which we have no evidence at all, such that our actual moral judgments might be, for all we know, wildly mistaken. Such an idea is a corollary of an ultra-realist conception of morality, and yet I suspect it is one at which even most so-called realists will balk. (Recall that this was one of the horns of Street's dilemma against the realist.)

Similarly, if we had some independent information about the probability of there being moral facts, then we might be able to support the stronger conclusion. Suppose we knew that moral facts were improbable, but took our moral judgments nevertheless to provide some support for their obtaining. The discovery that these moral judgments stem from a non-truth-tracking source would undermine this support, thus putting us back in the position of judging moral facts improbable. (This is not exactly *disbelief*, of course, yet framing the issue in Bayesian terms of degrees of belief is probably how the more nuanced presentation should proceed.[8]) Yet assessing the prior probability of moral facts obtaining is also a very difficult question regarding which there will be

[8] See Sober 2009b: 129.

nothing remotely like a consensus among philosophers (see Brosnan 2011: 55). So this route seems even less propitious for the error theorist than that sketched in the previous paragraph.

A more promising way of supporting the stronger premise 2 is via the endorsement of some methodological principle that underwrites it. Methodological empiricism, for example, will typically demand the *banishment* of any putative entity that fails to connect appropriately with perceptual input. Empiricism will often urge disbelief, rather than the withholding of belief, for any item that fails the test. (Recall Hume's directive that any book that doesn't pass empiricist muster must be "committed to the flames.") Even without specifying any particular version of empiricism, we can be confident that explanatory impotence will count as a failure, since such impotence implies an inability to figure in any perceptual process.

This last route seems to me the most plausible way of defending premise 2, though on the face of it seems rather dogmatic: Basically, one just embraces a methodological principle that demands (or at least permits) disbelief in explanatorily impotent entities. Presumably, though, the air of dogmatism may be dispelled by sensible considerations in favor of the methodology. The traditional school of empiricism, for example, wasn't based on a doctrinaire whim; its precepts were adopted for credible reasons. Whether premise 2 is plausible, then, will depend on an assessment of the considerations for and against the broader methodology that underwrites it.

Even if premise 2 is defensible, however, we have seen that there are many other "if"s in an error-theoretic debunking argument of this sort, and the argument strays a long way from the genealogy of morals with which it began. Ultimately, moral nativism may find a place as a premise in an error-theoretic debunking argument, but it will be a supporting role; the main actors will be propositions of a metaethical nature.[9]

Noncognitivist Debunking

The two styles of debunking argument thus far discussed—epistemological and error-theoretic—presuppose metaethical cognitivism: Moral judgments can be deemed epistemically unjustified or deemed false only if they are the kind of thing that can have truth value. Rejecting this presupposition, it would therefore seem, is a way of sidestepping the whole debunking dialectic. But perhaps a similar kind of debunking challenge can be devised for the noncognitivist.

Simon Blackburn's quasi-realist project takes an irrealist ontology, a noncognitivist construal of moral judgments (according to which they express conative attitudes), and from this basis endeavors to earn the right to the trappings of realism: talk of

[9] The tentative attitude expressed here toward a genealogical debunking argument in favor of a moral error theory must not be mistaken for a tentative attitude toward the conclusion. I stand by the error-theoretic metaethical position for which I have argued on other occasions (Joyce 2001, 2007b, 2011a); the question under current scrutiny is whether genealogical considerations can be used to establish that view.

beliefs, truth, assertions, facts, and so on (Blackburn 1984, 1993). It is difficult to integrate quasi-realism into many metaethical debates. Should it be assessed as an irrealist noncognitivist thesis, or as a position that supports moral truths, beliefs, properties, and so on? In the present context, what needs to be noted is that if the quasi-realist program succeeds in vindicating talk of moral properties, beliefs, and truths, then, to whatever extent the epistemological and error-theoretic debunking arguments work, they will apply to quasi-realist noncognitivism. I propose, then, to put the quasi-realist program to one side and work with a very simple and old-fashioned version of noncognitivism, according to which moral judgments as mental states are of a purely affective kind, and moral judgments as speech acts function solely to express those states. Let the states be simply some special form of *liking* and *disliking*. (I say this in order to exclude complications that would arise from treating noncognitivism as the view that moral judgments express *emotions*. The complication is that many emotions are mixtures of affective and cognitive components, and thus the epistemological or error-theoretic arguments could apply to the cognitive elements.) According to this view, there are no moral truths, no moral beliefs, no moral properties, no moral assertions, no moral knowledge.

Even if affective states cannot be false or epistemically unjustified, they can be mistaken in various other ways. Hume allows two ways for passions to be "contrary to reason": first, when based on a false belief about something's existence; second, when based on a false belief about what means are necessary and sufficient to satisfy some desire (Hume [1740] 1978: 416). It's not obvious what Hume means by "contrary to reason" in this context; he doesn't necessarily mean that these are the only two ways in which passions can be normatively appraised (see Schafer 2008). But even if he were to mean this, he is pretty clearly mistaken. If a passion is based on a belief that is not merely false but *irrational* (in the sense, perhaps, of being maintained irresponsibly in the face of discrediting evidence) then presumably the passion inherits a more serious kind of wrongness.[10] The taphephobe suffers from an irrational fear of being buried alive, but it is plausible that this fear is based on an irrational (and not merely false) belief concerning the likelihood of this occurring. Often phobic fears are irrational in another sense: because the fear is had in the absence of appropriate beliefs. An arachnophobe feels fear in the presence of a harmless spider, while knowing that it is harmless. I might like someone while believing (sincerely and truly) that she has all the qualities that I despise in a person and no redeeming features. Here it would be completely natural to assess my liking as "bizarre" and "irrational."

Hume will be quick to point out that in all these cases it is not the passion per se that is at fault, but rather that its error derives from its relation to belief: The passion either stands in the wrong relation to a good belief, or stands in the right relation to a faulty belief. "[P]assions can be contrary to reason only so far as they are *accompany'd* with some judgment or opinion" (Hume [1740] 1978: 416). But it seems that affective states

[10] And of course irrational beliefs are not a subset of false beliefs. An irrational belief may be true.

may also be subject to criticism without reference to beliefs. Consider our tendency to call imprudent attitudes "unreasonable." One's liking of something may cause one harm. (In such a case, one might dislike one's liking of the thing. Or one might not: One might like one's self-harming liking, which may well bring one further harm.)[11] It might be thought that imprudent passions are a special case of having false beliefs about the best means to satisfy one's desires—but this is plausible only on the assumption that people must always desire their own flourishing. Yet even when it is recognized that a person has sincere self-destructive devil-may-care desires, we do not cease to call his/her self-sabotaging actions and passions "imprudent."

Whether Hume allows this last category of evaluation isn't really my concern. That he does not is the tempting conclusion to draw from his memorable declaration: "'Tis as little contrary to reason to prefer even my own acknowledge'd lesser good to my greater, and have a more ardent affection for the former than the latter" ([1740] 1978: 416). But arguably even here Hume is making a claim about what the faculty of reason is capable of accomplishing, not placing a restriction on how imprudent preferences may be normatively assessed (see Schafer 2008). In any case, imprudent affective states *are* typically called "unreasonable" and "irrational," and the revelation that we are dealing with someone so aberrant as to consciously *prefer* his "lesser good to his greater" does not force us to retract the criticism. Perhaps there is nothing that could be said to such a person to change his mind; perhaps we'll go so far as to say that if these are really his preferences then he has no reason to refrain from pursuing the lesser good. Never mind; we can still legitimately criticize his preference as "unreasonable."

Not only are affective states subject to various kinds of criticism, but genealogical considerations will frequently form the basis for the criticism. Just as we cast doubt on someone's belief with the vernacular "You only believe that because…," so too do we disparage someone's attitudes by saying "You only feel that because…" A person's irritation may be dismissed by observing that she is tired. A person's preference for a musical performance is discounted on the ground that the performer is his daughter. A feeling of disgust will be convicted of some kind of misfiring if it is revealed that it was prompted by hypnosis.

It is not immediately evident precisely what these verbs of "dismissal," "discounting," and "being convicted of misfiring" denote.[12] Consider the last example just mentioned. Psychologists Thalia Wheatley and Jonathan Haidt (2005) hypnotized subjects to feel a pang of disgust upon hearing a given mundane word, like "often" or "take." The subjects were then presented with vignettes and asked to morally assess one of the characters therein, named "Dan." Those who had been hypnotized and given the trigger word were much more inclined to assess Dan negatively, even when no form of transgression

[11] An affective state may also be criticized for the harm it brings to others, but since this kind of moral assessment is the very topic that is under scrutiny, it is best put aside.

[12] Compare Daniel Kelly's comment on the use of the word "problematic" (Kelly 2014).

had been described. Upon being questioned, they confabulated grounds for their condemnation, or simply said things like "It just seems so weird and disgusting," or "I don't know [why it's wrong], it just is."

It is quite clear that we will dismiss disgust that has been prompted in this fashion. Perhaps we dismiss the disgust because of faulty beliefs that the subject holds: When she experiences disgust upon reading about Dan's actions, she might be inclined to "trust" her negative arousal to be a response to something disgust-worthy. And this is where she has gone wrong, because, unbeknownst to her, her negative arousal is definitely a response to a mundane word and not anything to do with Dan. Her belief that Dan's actions warrant disgust is mistaken, her belief that her emotions are a response to something that merits them is mistaken, and thus we dismiss her disgust and, further, dismiss the associated moral judgment (once we realize that the disgust is causally responsible for it). If noncognitivism is true, then the subject's moral judgment just is an expression of that disgust—or at least an expression of the dislike that the disgust prompts. The situation presented by Wheatley and Haidt would appear, then, to be a clear-cut case of genealogical debunking, even for a noncognitivist.

It might be useful to compare this with another case. Suppose you are slipped a pill that gives you a headache. It doesn't seem in this case that your headache, despite its unusual and secret origin, is (or can be) in any sense "dismissible." Generally one doesn't have very specific beliefs about the cause of a routine headache—perhaps a suspicion (a probabilistic belief?) that it's caused by dehydration. But in any case there is no belief about the headache's being *warranted* by its cause. Even when one firmly believes that dehydration has caused a headache, issues of *merit* just aren't apposite. Disgust is different in this respect. Disgust (unlike headaches) is an emotion, and this implies that disgust is more than just an affective state—it also involves or is accompanied by cognitive thoughts (for example, concerning affective states being merited by certain events). This, it would seem, is what makes the difference, allowing the emotion—including the affective component of the emotion—to be in some circumstances dismissed on grounds of its genealogy.

Talk of "dismissal" remains vague. I am certainly not saying anything about how we should treat a person whose emotions and moral judgments have been manipulated by, say, hypnosis. There may be various reasons for not pointing out to her what has happened. In the same way, were one to be a moral error theorist and hold that most people have false beliefs about morality, nothing obvious follows about how one should treat them. (Atheists generally don't go around knocking on theists' doors and teasing them.) The key point is that "dismissal" amounts to some kind of unmistakable albeit vague *undermining*—and this is sufficient for my present purposes.

Could the kind of unusual and local effect generated in the Wheatley and Haidt experiment scale up to a more ubiquitous debunking argument? Daniel Kelly thinks so (2011, 2014). He argues that the human disgust response evolved as an adaptive mechanism for dealing with the twin threats of toxins and parasites; this psychological response was then co-opted for negotiating social norms (which would explain the

connection between disgust and moral judgment revealed by Wheatley and Haidt). This genealogy forms the basis of a debunking argument:

> The emotion remains overly sensitive to cues related to its primary functions of protecting against poisons and parasites, which results in many false positives even in those domains. There is no reason to think the situation improves when disgust operates in the socio-moral domain. (2014: 134)

Kelly concludes that disgust "is not even remotely a reliable indicator of moral foul play… [and] feelings of disgust themselves should be given no weight in deciding whether an issue… is morally acceptable or morally problematic" (2011: 148).

Joshua Greene develops a similar argument (2008). Certain deontological moral intuitions, he argues, are driven by emotional mechanisms that played an adaptive role in our prehistoric past but which now fire in response to morally irrelevant factors. Faced with "trolley problem" scenarios, for example, subjects express reluctance to push a large person to his death off a footbridge in order to prevent a runaway trolley from killing five workers on the track, but are considerably less reluctant to save the five by pulling a lever to divert the trolley onto a side-track resulting in the death of a distant individual. The explanation, Greene argues, is that the former scenario triggers psychological mechanisms concerned with dealing with "up close and personal violence," the emotional effects of which generate "moral intuitions" against the former action but not against the latter action. The emotion-driven moral intuitions produced by the evolved human brain pertain not only to personal violence, but to retributive tendencies, to non-harmful actions (like food taboos), and to harming specific versus undetermined individuals. In Greene's opinion, these moral intuitions, coupled with a human tendency toward "*post hoc* confabulation," are responsible for deontological moral theory.

Greene uses these theses (for some of which he presents empirical evidence) as the basis of a debunking argument. "There are good reasons to think," he concludes, "that our distinctively deontological moral intuitions (here, the ones that conflict with consequentialism) reflect the influence of morally irrelevant factors and are therefore unlikely to track the moral truth" (2008: 69–70). Greene's target is larger than Kelly's, but is still selective; Greene believes that consequentialist moral thinking remains undebunked. (See also Singer 2005.)

One might be tempted to go further still, and aim for a *global* genealogical debunking of affect-based moral judgments. After all, the earlier debunking strategies within a cognitivist framework had global ambitions; why not also those within a noncognitivist framework? Suppose that the special forms of *liking* and *disliking* which I am assuming lie at the heart of noncognitivism are more like disgust than like headaches: that is, they are given practical weight because they are thought to provide insight into the like-worthy and dislike-worthy contours of the world. Just as disgust prompted by the word "take" is discredited if the person believes she is responding to something else (to someone else's transgressions, for example), so too would an act of liking be

discredited if it were discovered that the person is badly mistaken about what factors have aroused the state. Suppose that the mechanisms producing these liking and disliking responses can be given a particular kind of evolutionary explanation: They emerged because they helped bolster various cooperative motivations in our ancestors. This may reveal that we are ordinarily mistaken about what factors in the environment our affective states are responsive to: The states do not provide the touted insight into the like-worthy and dislike-worthy aspects of the world—they are not truth-tracking at all, but simply influence our motivations in ways that were once adaptive (perhaps via having a truth-tracking phenomenology). As with the case of hypnotically induced disgust, such false beliefs may be sufficient to discredit the affective states and thus the moral judgments that express them. The upshot may be nothing so radical as the prescription that we must attempt forthwith to purge our minds of these affective states (even if we could). The conclusion may be more analogous to the epistemological burden-of-proof-shifting discussed earlier: namely, that these affective states are left with a question mark hanging over them: They are not to be accorded the benefit of the doubt, they are not to be granted any privileged role in decision making.

Thus far I have had little to say about another obvious way of negatively evaluating affective attitudes: judging them detrimental to one's welfare. Even headaches can be assessed in this fashion. We tend to think of moral judgment in general as a prudentially good thing, but this is more of an item of faith than a properly scrutinized empirical thesis. Moral judgments can also be disastrous for those making them and for those around them. (Just think of all those patriotic young men who ended up as corpses in the trenches of the Great War.) A number of philosophers have pushed the view that on the whole we would be better off in practical terms if morality were eliminated from our mental and social lives (Hinckfuss 1987; Moeller 2009; Garner 2010; Marks 2013). It is not my task to evaluate the case(s) offered, but rather reflect briefly on how *genealogical* considerations might reveal the imprudence.

The argument follows a pattern by now growing familiar. Ordinarily, we might be willing to grant our affective states (like liking and disliking) the benefit of the doubt. We know we are evolved beings, and we might vaguely presuppose that evolution has designed us reasonably well. Pain exists to motivate us to respond to bodily injury, fear exists to motivate us to avoid danger, and so forth. Therefore when we feel pain (or fear, and so on), we have ground—at least in the absence of any reason to think otherwise—for assuming that its distinctive stimulus event is present, and that it is probably prudent for us to act as the pain (or fear, and so on) moves us to act. The same may be true of the affective states lying at the heart of noncognitivism (whatever they may be): We may take ourselves to have ground—in the absence of any reason to think otherwise—for assuming that it is probably prudent to allow these feelings a significant role in guiding our decisions. But this is precisely where a more detailed genealogical explanation can have an undermining impact, for it can reveal that the circumstances that rendered these affective states adaptive on the African savannah (say) no longer hold in the modern world, or fail to hold in some particular circumstances. Genealogical

evidence can act as a defeater of the benefit of the doubt we would otherwise accord an affective state—over-turning the assumption of its contribution to our welfare. Genealogical evidence can thus help reveal an affective state to be imprudent.

If the preceding arguments seem all rather slapdash, it is because my goal has not been to advocate them, but rather to highlight the fact that if these genealogical debunking arguments work at all, they work just as much against metaethical noncognitivism as against cognitivist success theory. Even if the noncognitivist is correct that moral judgments are no more than expressions of liking and disliking, these moral judgments can still be undermined by data concerning their evolutionary origins. Of course, this "undermining" won't amount to *being false* or *being epistemically unjustified*, but it cannot on this ground be dismissed as unimportant.

Conclusion

Genealogical debunking arguments are varied, not only in their premises but in their conclusions. They may or may not focus on the evolutionary perspective. Sometimes they rely on a principle of parsimony in the service of a radical ambition to establish an error theory; sometimes they attempt to shift the burden of proof in the service of a more modest epistemological conclusion. Although usually operative against the background of cognitivist presuppositions, genealogical debunking arguments can also have force within a noncognitivist framework. While there may be some convenience in lumping genealogical debunking strategies together as a family of philosophical arguments, in order to be effective any reasonable critic must discriminate among strategies and deploy counter-arguments applicable to his/her chosen target.

PART III

Projectivism and Fictionalism

9

Patterns of Objectification

John Mackie thinks that the "objective prescriptivity" with which our moral discourse is essentially but so fatally imbued is the result of our "tendency to read our feelings into their objects" (1977: 42). He invokes Hume's famous projectivist image of the human mind's "great propensity to spread itself on external objects," and, indeed, it is in his book-length analysis of Hume's moral theory (Mackie 1980) that the topic receives a more careful discussion than in *Ethics: Inventing Right and Wrong*. In both books he musters some considerations in favor of the objectification thesis, and reveals to us that he thinks that "it is very largely correct" (1980: 72).

But what is the relation between Mackie's objectification thesis and his thesis of moral skepticism? Is the error theory a premise in an argument to establish the objectification thesis? Or vice versa? Or are they logically unrelated? On the face of it, the objectification thesis appears to entail the error theory, but if this is so then one is left wondering why, in his 1977 book, this thesis is described straight *after* the argument from queerness, for if Mackie took himself to have some arguments in favor of moral objectification, then might it not have been strategically viable for him to establish the thesis of objectification *first* and then by implication argue for the moral error theory?

Wondering whether Mackie might have chosen to establish his moral error theory on the basis of the thesis of objectification is just my dialectical point of departure. My principal goal in this essay is to try to get a firmer handle on just what the thesis of objectification really is, and to investigate what evidence might support it and what conclusions may follow from it. I will disambiguate two forms of the thesis. One does trivially imply moral skepticism but cannot be established independently of that skepticism, whereas the other may well be substantiated on independent grounds but is neutral on the matter of moral skepticism.

First, some terminological clarifications. I will use "moral skepticism" and "moral error theory" interchangeably, reflecting Mackie's own practice. I will also interchangeably use the verbs "objectify" and "project"—and the associated nouns "objectification" and "projection," and "the thesis of moral objectification" and "moral projectivism." Mackie does this himself, in both his 1977 and 1980 books.[1] In this terminological vein, let us start by distinguishing "objectification" from some similar notions.

[1] The disjunctive phrase "projection or objectification" appears on p. 42 of Mackie (1977), and twice on p. 72 of Mackie (1980).

Hypostatization (also known as *reification*) is the practice of taking something abstract and speaking of it (or thinking of it, or treating it) as if it were concrete (for example, "Religion was his guiding light," "Justice is blind"). As a literary device or simply as a metaphorical manner of speaking, there is obviously nothing objectionable about the tendency; we do it constantly. If, however, one began to accept such propositions literally, at face value, then that would be a kind of error. Sometimes, though, it may be the subject of dispute whether an instance of this is an error. When philosophers try to provide a concrete explication of a seemingly abstract concept—like *set* or *number*—then some will treat this as an admirable extension of the naturalistic program while others will consider it a misguided blunder.

Anthropomorphism is often said to be a special type of hypostatization, wherein aspects of the inanimate world are imbued with human qualities. This taxonomy cannot be quite correct, however, since the subject of anthropomorphism need not be abstract. "Nature abhors a vacuum" attributes a human quality (abhorrence) to an abstract entity (nature), whereas "That stretch of road is treacherous" attributes a human quality (treachery) to a concrete entity (the road). As before, there is nothing wrong with anthropomorphism as a literary device or figure of speech, but one commits a kind of straightforward error if one really believes, of something incapable of human qualities, that it has such qualities. And, as before, there are areas of dispute, such as what kind of mental attributions can be made to animals or computers.

In *Modern Painters* of 1856, John Ruskin gave the name *the pathetic fallacy* (from "pathos") to a certain anthropomorphic tendency in writers and poets. He derided tired and uninspired anthropomorphic devices ("it is only the basest writer who cannot speak of the sea without talking of 'raging waves,' 'remorseless floods,' 'ravenous billows,' etc." (Ruskin [1856] 1908: 65)), but thought that some anthropomorphisms may be aesthetically justified when they reveal something genuine about the emotional life of the poet, despite being packaged with a false surface expression. (Still, Ruskin thought that the very best poets should be able to do away with such devices altogether.) Ruskin's anachronistic ruminations on aesthetics need not detain us; I mention the pathetic fallacy here because Mackie explicitly refers to it himself. He says that his thesis of moral objectification *is analogous to* the pathetic fallacy (1977: 42).

It is not obvious in what sense, exactly, the two are supposed to be analogous, but it is certainly important to observe a disanalogy. Witness what Mackie goes on to say immediately following, supposedly explaining the pathetic fallacy: "If a fungus, say, fills us with disgust, we may be inclined to ascribe to the fungus itself a non-natural quality of foulness" (1977: 42; he uses the same example in his 1946 article). This is actually not a good example of the pathetic fallacy (though it may be a good example of what Mackie is really driving at). The pathetic fallacy proper would occur if one ascribed to the fungus the human trait of *being disgusted*—which one would do, obviously, only in the context of joking or speaking metaphorically. By contrast, to attribute to the fungus the quality of foulness (irrespective of whether that property is non-natural or non-existent) is not a piece of anthropomorphism at all. In this example, a certain

feeling—disgust—has caused the person to experience the world as containing a certain quality—foulness. It seems pretty clear from context that Mackie will say that this foulness is not a real quality of the fungus; we think that it is only because our disgust somehow leads us to experience matters this way. Thus, to the extent that the fungus-foulness example exemplifies Mackie's notion of objectification (which I think it does), objectification is not an instance of anthropomorphism. Contrary to the quote given in the opening paragraph, the process of objectification is *not* one by which we "read our feelings into their objects"—we do not see the inanimate world as being disgusted, being angry, being sad (the waves as raging, the floods as ravenous), and so forth.[2] Rather, the process is one by which our feelings cause us to read into their objects qualities that we would not otherwise judge them to have—that (it is tempting to say) they do not really have at all, that nothing has. In his 1980 book, Mackie uses the vague but more accurate phrase: that we "read *some sort of image of* [our] sentiments" into the actions and characters that arouse them (71): the foulness is "some sort of image" of our disgust; the wrongness is "some sort of image" of our disapproval; and so on. In his much earlier 1946 article, he writes that "in objectifying our feelings we are also turning them inside out . . . The feeling and the supposed quality are related as a seal or stamp and its impression" (81–2).

Note that the characterization of objectification just given has several components. First, there's a complicated claim about our experience of the world and what has caused that experience (a claim I will attempt to make more precise in due course); second, there's a metaphysical claim that the quality we are experiencing the world to have is not really instantiated at all.

One might think that if this is objectification, then the error theory would indeed follow by implication. But that is not quite so, for the characterization just given does not exclude a *noncognitivist* interpretation of the matter. Let noncognitivism be the metaethical view that moral utterances are not assertions. The characterization of objectification says nothing, nor implies anything, about the nature of the speech acts that one might use to communicate one's experience. Suppose, for example, that we experience Xs as having the property Q, but in fact nothing instantiates Q—our experience is brought about by some emotional mental state E. This characterization is neutral regarding what is going on when we utter the sentence "X is Q." Perhaps we are asserting that X is Q (in which case an error theory holds) or perhaps we are merely expressing the emotion E (in which case noncognitivism holds). Thus, in order to have a characterization of objectification that implies an error theory, we would have to add a further clause: that our utterances about the subject matter in question are assertions (or that they are false, or that they are erroneous, or something else along such lines).

[2] Compare Barry Stroud's discussion of Humean projectivism: "We do not think that the sequence of events on the billiards table—the one ball's striking the others and the second ball's moving—itself has a feeling or impression like the feeling Hume says we humans get when we observe it . . . Nor do we think that an act of willful murder itself has a feeling of disgust or disapprobation, any more than we think that a painting on a wall has a sentiment of pleasure or awe" (Stroud 2000: 22).

It is possible that Mackie did have in mind such a complex account of objectification, though it is difficult to say with confidence. He several times states that the qualities that are projected "are fictitious" (1980: 71), but one cannot tell whether he intends this to be a defining component of objectification itself or just an additional point that follows from other things that he has argued. (I suspect the matter was not precisely demarcated in his thinking.) The point to which I should like to draw attention is this: If one does have in mind the complex account of moral objectification—the one that does imply the moral error theory—the one that consists of (i) a complicated claim about our experience of the world and what causes that experience, (ii) the metaphysical claim that the quality we are experiencing the world as having is not really present at all, and (iii) the cognitivist claim that our associated utterances are assertions—then whatever arguments one employed to establish this thesis would already be entirely sufficient to establish moral skepticism. The moral error-theoretic position, after all, just *is* the conjunction of components (ii) and (iii). Therefore, if we are using this complex account of objectification, then the argumentative strategy mentioned at the start of this essay—of establishing the moral error theory on the basis of the thesis of objectification—turns out not to be dialectically viable, after all.

Suppose, alternatively, that we worked with a less complicated version of objectification. Suppose we stripped away both the cognitivist sub-thesis (iii) and the metaphysical claim (ii), leaving just (i) a claim about our experience of the world and what causes that experience. Elsewhere (Joyce 2009a) I have called this remaining theory "minimal projectivism"; here, in line with Mackie's preferred terminology, let us call it "minimal objectification." Would it be viable to argue for moral skepticism on the basis of minimal objectification? In order for a positive answer, minimal objectification must imply moral skepticism, but it must do so non-trivially—it cannot be that making the case for minimal objectification would require first making the case for moral skepticism.

Before proceeding, we must be more precise regarding this "claim about our experience of the world and what causes that experience." This claim is the heart of objectification; it is the part that is supposed to capture, in literal terms, the whole idea of something mental being *projected* onto the world. It is not my intention here to offer a *general* account of objectification, for that poses a number of complications that are surprisingly challenging to overcome and the effort is not necessary on this occasion. I am satisfied to sketch an account of a pertinent proper subset of objectification: minimal *affective objectification*.[3]

[3] For the curious, I will quickly outline the difficulties of providing a general account of minimal objectification (or "projectivism" as I will call it in this footnote, since I will relate it to subjects for which the latter term is more familiar). An adequate general account of projectivism should cover at least the following cases: (1) moral projectivism, (2) causal projectivism, (3) psychopathological projectivism, and (4) color projectivism. In the interests of illustrating the difficulties of achieving a general account, let me sketch, in the most provisional terms, what these four theories might look like. In moral projectivism, something in the world prompts one to feel disapproval (say), which leads one to experience the thing in the world as forbidden. In causal projectivism, a regularity in the world prompts one to have an expectation, which leads one to experience the world as containing a causal relation. I am not confident that I can

S's experience of X as P is an instance of minimal affective objectification if and only if: (1) S experiences P as an objective feature of X, and (2) this experience has its immediate causal origin in some affective attitude (for example, the emotion of disapproval) rather than a perceptual faculty.

Call (1) the *phenomenological thesis* and (2) the *causal thesis*.[4]

We experience the fungus as objectively foul, but this experience is the immediate result not of perceiving the property of foulness, but rather of our emotion of disgust. Naturally, perception is involved prior to our feeling disgust: We *see* the fungus, we *smell* the fungus, and so forth. It is in acknowledgement of this prior engagement of perceptual faculties that the causal thesis claims that the experience (of foulness, for example) has its *immediate* origin in an affective attitude rather than a perceptual faculty. (That this is a vague and potentially problematic qualification does not escape me.) This appearance of "immediate" is designed also to exclude non-projection cases where some affective state, such as emotional arousal, guides subsequent perceptual processes, such as focusing one's attention on aspects of the world that might otherwise escape notice. An intensely jealous person, for example, might notice something in the body language of another that she otherwise would not have seen. Let us say that the body language is real: The other person really is (say) sitting with uncrossed legs. And the jealous party experiences this body language as an objective aspect of the world (thus satisfying the phenomenological thesis). And were she not jealous, she would not have noticed it. However, her experience is not an instance of minimal affective projection because it does not flow *immediately* from the affective attitude of jealousy; rather, the jealousy has guided the subsequent engagement of her perceptual faculties. In the case of foul fungus, by comparison, the disgust does not lead one to see the fungus as foul via channeling one's perceptual attention to aspects of the fungus (though it may incidentally direct perceptual attention as well); the relation between the disgust and the foulness—the relation that Mackie described as the latter being "some sort of image of" the former—is more direct than that.

give a general account of all forms of psychopathological projectivism, but examples are not hard to come by: A person's poor self-image leads her to interpret her parents as being overly critical or demanding. So far we have three mental states that serve as "intermediaries" between the world and one's experience of the world: disapproval, expectation, and a poor self-image. Yet it is hard to come up with an over-arching category for these three that will not end up capturing too much. And in the case of color projectivism, it is challenging even to come up with an analogous intermediary mental state. The idea is (roughly) that one's visual experience of color owes its quality to the nature of one's sensory apparatus rather than to the real nature of objects' surfaces (even though the sensory apparatus is sensitive to real properties of surfaces). It is not obvious what mental activity is supposed to be getting projected in the creation of one's visual color field. (Visual qualia?) My hunch is that color projectivism can be articulated only as a metaphysical (non-minimal) thesis. It is in light of these kinds of complications that I have sidestepped the delicate task of trying to unify this family of stock examples (and others besides) with a general account, though I confess to harboring the hope of yet doing so in the future. Perhaps in the end there is no entirely satisfactory general account of minimal projection in the offing.

[4] See Joyce (2009a), where some features of this account (for example, what might be meant by "objective") are discussed in more detail.

In order to articulate the thesis of minimal *moral* affective objectification we need do nothing more than stipulate that "P" (as in "… experiences X as P") stands for a moral adjective. I cannot think of a theory of moral projectivism for which the mental state putatively "projected" in the creation of moral experience is supposed to be anything other than an affective attitude (for example, disapproval, subscription to a normative framework, and so on), so in what follows I shall drop the "affective" qualification. Clarity requires, however, that we keep the "minimal." I do not suppose that every philosopher who has spoken of "moral projectivism" or "moral objectification" over the years has really had in mind, even tacitly, the minimal variety.

Both the phenomenological thesis and the causal thesis appear to be psychological claims. Quite how we would go about empirically testing them for the case of morality (or any other case) is a nice question (into which I won't delve on this occasion), but it seems pretty clear that, on the face of it, whatever methods of empirical psychology we employed would not require us first to establish that moral properties do not exist, or that moral judgments are uttered with assertoric force. Therefore one desideratum of the strategy of arguing for moral skepticism on the basis of minimal objectification appears to be satisfied: The latter does not imply the former *trivially*.

The other desideratum, however, is not satisfied. Minimal moral objectification does not imply moral skepticism *at all*, for it is metaethically neutral. Not only is it silent on the cognitivist/noncognitivist debate, it is also silent on the metaphysical debate over the existence or non-existence of moral properties. It is, therefore, compatible with moral realism and thus does not imply moral skepticism. Let me explain. Note, first, that there is a *prima facie* pressure in favor of maintaining this compatibility, based simply on the observation that were the phenomenological and causal theses to be confirmed by empirical inquiry (as they very well might be), it would surely be astonishing if the moral realist were to roll over and concede the game to the skeptic. It may nevertheless clarify matters to explore this compatibility in more specific terms. A simple example will suffice to get the compatibility on the table, and then I will develop a more satisfactory example for the moral case.

Consider an everyday usage of the notion of *projection*. Suppose that a person tends to experience others she encounters in social situations as critical and reproachful; but in fact this is due to her own meek and self-doubting nature. It would not be at all peculiar for us to describe this person as *projecting* her sense of her own inadequacy onto others. Now suppose that on a given occasion this person encounters a man who really is unusually critical and reproachful. Let it be stipulated that he has not yet indicated to the woman, even in the most subtle fashion, that he has such a personality; she simply assumes that he does, just as she does with everyone else that she meets. The woman's judgment "This man is critical and reproachful of me" is just as much a projection on this occasion as ever it is; and this conclusion is not undermined by the fact that the judgment happens this time to be true. Even if the woman *often* encountered critical and reproachful persons, her subsequent negative judgments about them (which will now *often* be true) might nevertheless remain instances of her projecting

her sense of inadequacy, so long as we specify that she would have formed these same judgments even if these critical and reproachful persons' personalities had been otherwise.

An analogous situation in the case of moral judgment would suffice to show the compatibility of minimal moral objectification and moral realism. We can imagine a scenario in which, on the one hand, moral judgments are acts of minimal objectification/projection, while, on the other hand, these judgments (when uttered) are assertions that are often true. According to one influential view, satisfying the latter conditions suffices for moral realism (Sayre-McCord 1986). Some would prefer to add a further clause to moral realism: that the assertions in question are true in virtue of some objective state of affairs. (See Joyce 2007a for discussion of the definition of "moral realism.") Let it be so; imagining the scenario in such a way that this additional clause is also satisfied will not spoil the example by undermining the projectivist stipulation.

But one may remain unsatisfied with this demonstration of compatibility, for the example had the projection-derived judgments turning out true *by coincidence*. The example shows minimal objectification and moral realism to be compatible according to the letter of the law, but perhaps not the spirit. Can we eliminate this aspect of accidentality from the demonstration of compatibility?

I believe that we can. Return to Mackie's example of foul fungus. Assume the minimal affective objectification account holds true: A person experiences the fungus as having an objective quality of foulness, and this experience has its immediate origin in the person's disgust. Now let us see whether realism about *foulness* might also hold true. There is certainly nothing to exclude us holding that when the person makes public her judgment via the utterance "That fungus is foul!" she is making an assertion. It is the other realist elements that might be deemed troublesome: that the assertion (a) is true, (b) is true in virtue of an objective fact, and (c) is non-accidentally true: that is, is made in a way dependent on the truth-rendering fact. (Whether (b) is really a necessary aspect of realism is moot, and that (c) is a necessary part of realism seems doubtful, but let us add these components for the sake of argument.)

Consider a response-dependent account of the property of foulness. (See Casati and Tappolet 1998; Johnston 1989, 1992, 1993; Wright 1988.) Foulness, on this account, is a disposition to produce a certain kind of psychological response R in a certain kind of subject S in a certain kind of circumstance C. It would be no challenge to specify these variables R, S, and C in such a way that certain items in the actual world—such as bits of fungus—have this disposition, thus rendering assertions of the form "X is foul" true. The account can also claim to satisfy an important kind of objectivity. (See Pettit 1991.) The disposition in question will be understood in modal terms, such that the fungus *would* prompt R if apprehended by S in C. Thus the fungus may have the disposition even if no Ss have ever been in C, even if no Ss have ever experienced R, even if no Ss ever have existed or ever will exist. Thus the disposition does not depend on the *existence* of any particular psychological state, or even the existence of psychological states

generally. (One might point out that the disposition is nevertheless *conceptually* mind-dependent—since it cannot be adequately described without reference being made to psychological state R—but it is not at all obvious that realism requires "conceptual mind-independence." It certainly doesn't require every form of mind-independence. One would like to be a realist about domesticated dogs, despite the fact that they are *causally* mind-dependent entities: existing only because of generations of *intentional* behavior on the part of humans.)

In order to emphasize the fact that satisfying these realist conditions has not undermined the projectivist assumptions that we started with, let us specify R in such a way that the psychological response in question is in fact an act of objectification. In other words, suppose that foulness is the disposition *to prompt Ss (when in C) to feel disgust and to objectify this disgust in their experience of the world*. This appears to build the satisfaction of the minimal objectification thesis (that is, the satisfaction of the phenomenological and the causal theses) into the response-dependent account. We are thus in a position to see that the judgment that is derived from a process of objectification—"That fungus is foul!"—does indeed manifest a dependence relation on the truth-supplying fact; it is not merely coincidentally true: Were the fungus not to have been foul, the observer would not have made the judgment.[5]

One might worry that in striving to satisfy the dependence relation the account has undermined the causal thesis. If there is this dependence relation in place—if the counterfactual that ends the last paragraph holds true—then isn't the person's experience of the foulness (the foulness that we are here supposing to objectively exist) a case of *perception*, after all? And if so, then the causal thesis is undermined, in which case my attempt to show the compatibility of minimal moral objectification and realism will have failed.

I respond not by offering a full account of perception (which I have neither the space nor the expertise to do), but by observing that the counterfactual dependence mentioned is certainly not a sufficient condition for perception. A couple of simple examples will suffice. Consider learning something from reading a book: Suppose one learns for the first time that Napoleon lost the Battle of Waterloo. The book, we are supposing, is an accurate one, in the sense that had Napoleon not lost Waterloo, the book would not have contained the claim that he did lose. And we will also suppose that the reader judges that Napoleon lost Waterloo solely on the basis of reading this reliable history

[5] Needless to say, Mackie himself won't buy the response-dependent account of moral properties. He criticizes such views in general terms in his 1980 book (chapter 5); and in his 1973 book (chapter 4) he doubts even the existence of dispositional properties. It is also doubtful that Hume's multifarious uses of the projectivist metaphor are supposed to be compatible with realism. Stroud (2000) emphasizes how, in Hume's account, the content of projectivist experience—be it causal connection, beauty, color, or virtue—is something that could not even be *intelligibly* predicated of items in the world. Immediately following the famous *Treatise* projectivist image of the mind's "great propensity to spread itself on external objects," Hume declares that sounds and smells "really exist no where" (Hume [1740] 1978: 167)—and context makes it reasonable to think that he will say the same of color and necessary connection. I discuss the error-theoretic commitments lying behind Hume's views in Joyce (2010b).

book. This judgment thus manifests the appropriate counterfactual relationship with the relevant fact, but we would not on this basis conclude that the reader *perceives* Napoleon losing Waterloo. The second example concerns someone judging that a certain object instantiates a dispositional property. Suppose someone works out that a vase is fragile by some means other than breaking it. Perhaps she smashes a lot of similar vases, or perhaps she asks some authoritative people. We will suppose that whatever means she employs is reliable, in the sense that had the vase not had the disposition, she would not have come to make the judgment about its fragility that she does make. But, though her true judgment ("The vase is fragile") manifests a dependence relation on the truth-rendering fact (the vase's being fragile), we would not say that she has literally *perceived* this vase's fragility.

Thus, in striving to come up with a realistic account of foulness that satisfies the desideratum of the judgment being dependent on the truth-supplying fact (along with satisfying all the other realist criteria, too), we have not undercut our starting projectivist assumption that the experience has its immediate origin in an affective attitude *rather than a perceptual faculty*. And what goes for foulness here can go for moral qualities, too. I conclude, therefore, that minimal moral objectification is compatible with moral realism—even a fairly robust version of moral realism.[6]

Whatever may seem surprising or counter-intuitive about this conclusion probably stems from the fact that *non*-minimal versions of objectification are more familiar to us, both in vernacular settings and in the philosophy classroom. We are more likely to describe the case of psychopathological projection in terms such as "Her sense of her own inadequacy makes her see others as overly critical *when really they're not*."[7] We are more likely to describe moral objectification in terms such as "Our feelings of disapproval and aversion lead us to see the world as containing moral qualities *that it does not really contain*." It is not my intention to condemn such non-minimal, metaphysically committed uses of objectification; they may, indeed, be the more natural and useful formulations in most circumstances. My point has been, rather, to show that in order to establish the truth of any such metaphysically committed objectification thesis one would need to have *already* shown that the quality in question does not exist, and so the strategy of supporting moral skepticism by means of first establishing a metaphysically committed version of moral objectification is unworkable. We have also seen that the strategy of supporting moral skepticism solely on the basis of establishing a *minimal* version of moral objectification is also unworkable, for the minimal version is silent on the debate between the moral realist and the moral skeptic.

As to this last claim, one might object that *of course* minimal objectification should be silent about these metaethical claims, for the whole point, after all, is that it should not imply moral skepticism *trivially*; it should imply it only in conjunction with some

[6] There is precedence for seeing projectivism and realism as compatible in the Humean literature. See Craig (2000) and Sainsbury (1998).

[7] Freud, remember, categorized this sort of projection as a kind of *delusion*—indicating an anti-realist construal.

other premises. Minimal objectification may well be consistent with moral realism, but perhaps these other premises will serve to narrow the range of possibilities so as to exclude realism along with every other theoretical possibility bar the error theory. One or more of these additional premises may function to knock noncognitivism out of the running. (Mackie does, recall, develop several arguments against the noncognitivist in both his 1977 and 1980 books.) Other additional premises may serve to establish the non-existence of moral properties. Perhaps when we add up all of these premises we end up with an argumentative route from minimal objectification to the moral error theory.

At first blush, the problem with this objection appears obvious. The additional premises adverted to would appear to be *sufficient* to establish the moral error theory; they would be doing *all* the work, and the thesis of minimal objectification would be entirely superfluous in this argument. Therefore, we seem to see once more that the strategy of establishing the moral error theory on the basis of the thesis of minimal objectification turns out not to be viable.

On closer inspection, however, the objection has more merit. Consider the passage with which Mackie introduces the topic of objectification, immediately after presenting his arguments from relativity and queerness:

> Considerations of these kinds suggest that it is in the end less paradoxical to reject than to retain the common-sense belief in the objectivity of moral values, provided that we can explain how this belief, if it is false, has become established and is so resistant to criticisms. (1977: 42)

The thesis of objectification is supposed to satisfy this "proviso." In other words, Mackie has exposed a theoretical option: We can either "reject . . . the common-sense belief in the objectivity of moral values" or we can "retain" it. The thesis of objectification is evidently supposed to function as the tie-breaker, making rejection the reasonable choice. We are forced to conclude that (Mackie thinks that) without the thesis of objectification, retaining the common-sense belief would remain a live option. Objectification functions as an explanation of where the massive error embodied in morality comes from, in such a way that *without that explanation there remains doubt that it is an error at all*.

There is a tempting alternative reading of Mackie, which accords the thesis of objectification a lesser role. According to this alternative, by the end of the section in which he presents the argument from queerness (section 9, chapter 1) Mackie has established the moral error theory to his own satisfaction. A reader might at that point accept the moral error theory but then be naturally *curious* to know where this widespread systematic human error has come from, and the thesis of moral objectification is supposed to satisfy this curiosity. According to this reading, the thesis of moral objectification does not function to satisfy a proviso, but is, rather, a supplement to the completed skeptical argument: an interesting explanation but strictly dispensable in establishing the case for moral skepticism. The above-quoted passage, however, indicates that this alternative reading is incorrect. The case for moral skepticism (embodied

in the arguments from relativity and queerness) is plainly considered incomplete until the thesis of objectification is put forward. But what role precisely does the thesis of objectification play in Mackie's case for moral skepticism?

The answer, I believe, lies in Mackie's almost tacit acceptance of some kind of epistemological conservatism. Just prior to presenting the arguments from relativity and queerness, he has admitted that since moral skepticism "goes against assumptions ingrained in our thought and built into some of the ways in which language is used, since it conflicts with what is sometimes called common sense, it needs very solid support" (1977: 35). Mackie appears to be acknowledging that the counter-intuitiveness of moral skepticism in itself represents a burden of proof that the error theorist must strive to overcome—that the error theorist's arguments need to be even more convincing than those of his opponent if he is to win the day.

That Mackie is an epistemological conservative may seem an unexpected conclusion, considering how accustomed we have grown to seeing the moral realist reach for the principle of epistemological conservatism as one the main weapons in the anti-skepticism arsenal. (See, for example, Huemer 2005; Brink 1989: 23–4; Dancy 1986: 172.[8] See Loeb 2010 for discussion.) But it is apparent that Mackie does indeed consider the "intuitiveness" of a philosophical thesis a valid consideration in deciding whether to endorse it. All else being equal, an intuitive theory is to be preferred over a counter-intuitive one; in other words, the very fact that a belief is held supplies it with a certain *prima facie* epistemological justification. And that there *are* intuitions in favor of morality is hardly to be denied; any error theory worth arguing about is, *ex hypothesi*, counter-intuitive.

Mackie's arguments in favor of moral skepticism, then, must overcome these standing intuitions. The vital role of the thesis of objectification is to *explain away* the content of these pro-morality intuitions by providing an account of their origin that does not imply or presuppose their truth. Such a genealogical explanation serves to defeat or block whatever *prima facie* justification these intuitions might otherwise have been granted. The skeptic does not deny or doubt the principle of conservatism; he takes it seriously. In particular, the skeptic attends to the principle's "all else being equal" clause. "When are things *not* equal?" the skeptic wonders. When are intuitions defeated, and under what conditions might they not even be accorded *prima facie* epistemic status? One answer (among many, no doubt) is that things are not equal if one has a plausible, or even empirically confirmed, theory of where the intuitions in question come from that is consistent with their being false.

The moral skeptic, then, needs two lines of reasoning. The first attempts to show that there is something fishy about moral facts. (For Mackie, this is achieved by the conjunction of the arguments from relativity and queerness.) But the conclusion of this reasoning is, all parties agree, counter-intuitive. A principle of epistemological

[8] I offer some criticisms of this line of argument—and of Michael Huemer's version of it in particular—in Joyce (2009c).

conservatism threatens to allow our pro-morality intuitions to override the skeptic's case. So the skeptic offers a second body of evidence: explaining away the content of those pro-morality intuitions by revealing them to be the product of an unreliable process of formation. The second move aims to show "how even if there were no such objective values people not only might have come to suppose that there are but also persist firmly in that belief" (1977: 49). The skeptic thus discharges the burden of proof with which he is lumbered—not by bolstering his initial argument (presenting more evidence in support of premises, devising new argumentative moves, and so on), but by casting into a doubtful light those very intuitions that promised to give the conservative principle traction. (Moreover, nor can these same pro-morality intuitions be raised as a consideration against the argument for objectification, since it is a corollary of the objectification thesis that such intuitions will be in place.)

My claim that the intent of Mackie's two skeptical arguments (relativity and queerness) is merely to show that there is something "fishy" about morality might itself seem fishy; surely, one might object, these arguments are supposed to be more decisive than that. If these are sound arguments (the objection continues), then they need no extra argument to act as a tie-breaker; and if they are not sound arguments, then why should we pay them any attention? However, I think this dilemma does not succeed in undermining the interpretation being offered. Let us briefly consider Mackie's two skeptical arguments in turn.

The argument from relativity takes the form of a competition between two hypotheses: The phenomenon of moral disagreement may be explained either by the supposition that some parties have privileged epistemic access to the realm of moral facts (the realist's hypothesis) or by the supposition that there are no moral facts at all (the skeptic's hypothesis). Among the many criteria that we might employ in deciding between these two hypotheses, a comparison of their levels of mesh with our intuitions may well figure. Certainly the epistemological conservative allows it to figure. Therefore putting forward evidence (such as the thesis of objectification) that *explains away* the content of intuitions in favor of one hypothesis is both strategically permissible and potentially determinative.

The same point is slightly less obvious in the case of the argument from queerness, for here, it might seem, we have an argument that purports to stand soundly on its own: Premise 1 is a piece of conceptual analysis (that moral discourse is centrally committed to the existence of objective prescriptions) and premise 2 is an ontological claim (that there exist no objective prescriptions). However, it is, I think, slightly naive to suppose that Mackie considers the argument from queerness to be a sound argument with demonstrably true premises. It is more realistic to think of it as providing a firm consideration in favor of moral skepticism, its premises having the status of hypotheses on whose acceptability many factors may have a bearing. There may, for example, be considerations to be raised in support of the claim that there exist no objective prescriptions, but also considerations to be raised in support of the claim that there do exist objective prescriptions. No impartial spectator who has been paying attention to the

debate could seriously doubt this; it is a complex and nuanced discussion that leaves plenty of wriggle room for competing interpretations of key elements (as some of the contributions to this volume[9] demonstrate). The epistemological conservative allows that if there are standing intuitions in favor of the view that objective prescriptions exist (as it seems reasonable to suspect), then these may be accorded a role in weighting the debate against premise 2 and thus against the argument from queerness. Therefore, again, the strategy of raising evidence (such as the thesis of objectification) that casts into doubt those very intuitions that speak against the argument from queerness—showing that they arise from a potentially unreliable source—is entirely legitimate and may very well swing the argument the skeptic's way.

We have seen, then, that the two lines of reasoning need each other. The case for moral skepticism is not achieved by the arguments from relativity and queerness alone; evidence to explain away the counter-intuitiveness of the conclusion (or individual premises) is also required in order to overcome the challenge from epistemological conservatism. And the thesis of objectification alone will not provide a skeptical conclusion. The minimal version of the thesis is metaethically neutral, and to employ a metaphysically committed version to this end would simply beg the question.

Thus far I have been concerned entirely with the role that the objectification thesis plays in Mackie's overall strategy for establishing moral skepticism, but we have not yet examined any of the arguments he provides to convince us that the thesis is actually plausible. I will close by running through the considerations in support of the objectification thesis that Mackie offers in his 1977 and 1980 books.[10]

In the interests of clarifying the structure of the argument we should first acknowledge that it was open to Mackie to eschew the task of mustering evidence in favor of the objectification thesis, and instead simply present the thesis as a coherent and possible hypothesis of the genealogy of moral judgment according to which these judgments are not, or might not be, true. But it is evident that this is not Mackie's attitude toward the thesis—and it is as well that it is not, for this strategy would place the objectification thesis in the same category as a host of other skeptical hypotheses that lack any real plausibility but which have the (dubious) virtue of thwarting all attempts at falsification. (Brains in vats and deceiving demons spring to mind.) The objectification thesis plainly isn't supposed to function in this disappointing way—merely as a skeptical hypothesis that *might*, for all we can prove to the contrary, hold true. It is supposed, rather, to be a serious hypothesis for which we can marshal solid evidence.

Nevertheless, in my opinion, Mackie misses much of the opportunity to establish the plausibility of the thesis by looking for supportive evidence in the wrong places. The few pages devoted to this in his 1977 book (43–8) are uncharacteristically somewhat obscure. In particular, he does not do an adequate job, in my opinion, of

[9] That is, Joyce and Kirchin (2010).

[10] Space does not permit an examination of Mackie's arguments for moral objectification found in his 1946 paper. I do not think any of the arguments found there are superior to those problematic ones which I shall discuss.

distinguishing between the thesis that moral experience is the result of our having objectified affective attitudes (feelings, wants, and demands are mentioned by him) and the thesis that we simply have false beliefs about the objective status of moral properties. That we should recognize such a distinction seems highly desirable. Consider something that we would ordinarily consider a non-objective matter—say, what counts as polite behavior at meal times. A person might foolishly believe that the propriety of keeping one's elbows off the table is an utterly mind-independent affair, and that all those other cultures that allow alternative rules of etiquette are simply mistaken. This person has a false belief about the objectivity of something. But must it be an instance of *objectification*? It might be, of course, but it also seems reasonable to suppose that it might not be. Objectification essentially involves a certain sort of psychological *operation* that leads to belief (or, speaking more carefully, that leads to a certain quality of experience). "Objectify" is a transitive verb; there must be something that gets objectified. But one can have a (false) belief about the objectivity of something without any operation of objectification having occurred. Mackie's failure to attend to this distinction undermines the force of many of the considerations he raises in support of the objectification thesis.

For example, Mackie begins by pointing out that a widespread belief in objective moral properties might fulfill certain human needs: Such properties (if they existed) would have a kind of practical authority over human affairs such that a widespread belief in their instantiation would regulate interpersonal relations in an effective way. This is a complex but broadly plausible claim. Yet it doesn't obviously provide any evidence in favor of the objectification thesis as opposed to the "false-beliefs-about-objectivity" thesis. And even as evidence for the latter it is weak: The consideration merely shows that we might have a motive for believing in objective moral properties, which falls short of demonstrating that we *do* so believe. One might try to wring from this some (proportionally weak) support for the objectification thesis by pointing out that if we have a *motive* for believing in an objective morality, and this motive does lead us to belief, then we have *objectified* that motive, thus satisfying the criteria of the objectification thesis. But this is a problematic line of argument. What it is to "have a motive" is a complicated and indeterminate matter. On one reading it means that there is a reason to do something, even if one is unaware of this fact. ("The Romans had a motive for ceasing to line their aqueducts with lead.") A more robust reading requires making reference to an agent's desire. ("Romeo had a motive for climbing to Juliet's balcony.") This indeterminacy creates a fatal dilemma for the argument under consideration.

Suppose, first, that when we say that humans "have a motive" for believing in an objective morality, we are using "motive" in something like the former sense, to mean that, as a matter of fact, things will go better for us (each of us, let's say) if we all have this belief. This, however, may be true while we all remain utterly ignorant of the fact, showing that such a motive might not exist in our psychological profile in any sense, and therefore is not *there* to be objectified. The item which we are calling "the motive" might nevertheless have causal powers. That fact that things go better for organisms with X

than organisms without X can lead to change—through a process of natural selection, for example. I admitted that it is broadly plausible that things will go better for groups of humans if they believe in an objective morality, and we can thus imagine this belief becoming fixed in a population through some process of cultural (or biological) evolution. All this would make it permissible to claim "Humans have a motive for believing in an objective morality, and this motive brings it about that they do so believe." But at no point need we maintain that the "motive" in question is psychologically represented by/to humans—let alone as an affective attitude—and therefore the criteria for objectification are not satisfied.

Alternatively, we could understand "motive" in the latter sense, to mean that humans (generally) do desire to believe in an objective morality, and this desire brings it about that they do so believe. This would satisfy the criteria of the objectification thesis, but only at the cost of being a fantastic empirical claim. Bearing in mind that we are using "desire" in the sense of *occurrent, affective attitude* (something that has causal powers within an individual's psychological economy), it is wildly implausible that humans typically desire to believe that moral properties are objective.

I conclude, therefore, that while it might be true that humans "have a motive" for believing in an objective morality, and that this fact might have had (and continue to have) an important causal role in bringing it about that we do so believe, this does not provide grounds for claiming that this belief is the product of our having *objectified* that motive.

The next strategy explored by Mackie (1977: 43–4) in seeking support for the objectification thesis is to give some specific examples of moral objectification, which he designates "patterns of objectification." Sometimes we desire something for perfectly sound (non-moral) reasons, but then we "confuse" this basis for our desire (the item's "subjective value") with the idea that the item in question has objective value. Sometimes we think that someone ought to do something for instrumental reasons, but then we suppress the instrumental conditional clause and claim that she ought do it *simpliciter*. We might do this because expressing it in this way is more likely to lead to compliance. Later (1977: 47), another kind of confusion is mentioned: when we muddle a descriptive and objective sense of the goal of humans (as in what we in fact pursue, or what posited goal will confer sense upon our actual actions) with a normative but subjective sense of that goal (as in what we *ought* to be pursuing).

These might very well be cases of objectification, and it is not at all implausible to suppose that Mackie is accurately describing some real human phenomena. But if his goal is to provide evidence that human moral judgment is *typically* (always?) the product of a process of objectification, then these examples hardly count as strong evidence. After all, I doubt that anyone (apart from some philosophers who worry in their characteristic manner about the details of the thesis of projectivism) will deny that moral judgment is *sometimes* the product of projected desires, emotions, and moods. Therefore, exposing some cases of moral objectification does not suffice to explain how the (putatively) false intuitions in favor of morality have "become established" and are

"so resistant to criticisms" (1977: 42). By analogy, suppose one were to doubt the existence of human character traits, and sought to explain away those strong intuitions we seem to have that such traits do indeed exist. One potential explanation would be the projectivist one: Our tendency to see others as instantiating stable character traits is the result of our projecting aspects of our own mental lives onto them. In providing evidence for this projectivist thesis, it would clearly not suffice to point out some instances of this kind of projectivism—such as the example I employed earlier of someone's sense of her own inadequacy leading her to see others as critical and reproachful. All parties (we can suppose) accept the existence of the phenomenon of psychopathological projection, but much more evidence would be needed to show that this phenomenon somehow generalizes in unexpected ways. By referring in the plural to "patterns of objectification," Mackie might hope to give the impression of a systematic and widespread tendency here (as in "There is a pattern emerging!"), but in reality he has not provided evidence sufficient for his purposes.

A further strand in Mackie's case for the objectification thesis is broadly historical. Once upon a time a lot of our moral language was embedded in an ontological framework that included an all-powerful, all-seeing, loving deity doling out rewards and punishments. Mackie concedes that if "this theological doctrine could be defended, a kind of objective ethical prescriptivity could be introduced" (1977: 48).[11] He looks with some sympathy upon Elizabeth Anscombe's conjecture that "modern moral philosophy" (*à la* mid-twentieth century) consists of trying to make sense of a family of normative concepts "outside the framework of thought that made [them] really intelligible" (Anscombe 1958: 6). However plausible this "conceptual residue theory" may be (and I agree with Mackie that there is surely something to it), it nevertheless seems to have little to do with the psychological process of objectification. Anscombe may have explained the origin of a widespread but false belief in the objectivity of morality (and therefore does provide resources upon which Mackie can draw), but the hypothesis does not fit with the desired model of this belief being the result of our having "spread" our wants and demands onto "external objects." I conclude, then, that the case for the objectification thesis in Mackie's 1977 book is very weak.

We find a more structured and clear argument for the objectification thesis in 1980's *Hume's Moral Theory*. Here Mackie claims that it is the only theory that can properly make sense of three phenomena: (i) that the evidence seems to favor metaethical cognitivism, (ii) that moral statements are taken to be intrinsically action-guiding (that is, not to rely on subjects' ends and goals for their legitimacy), and (iii) that "the essential fact of the matter, which underlies moral judgments . . . is that people have various sentiments" (1980: 72). Although the last is stated rather clumsily, it is apparent to what Mackie is referring. In the preceding chapter he has identified nine arguments used by

[11] Incidentally, this comment reveals that Mackie believes that so long as there exists some possible world at which the requisite kind of supernatural being is real, then the error in morality is but a contingent matter. In his forceful case for atheism in his 1982 book, Mackie repeatedly declares the existence of God to be "improbable" (100, 130, 252–3)—not impossible.

Hume to press the negative view that moral judgments are not the product of reason, and then he has provided a number of well-known passages from Hume variously stating the positive "plain hypothesis": "that morality is determined by sentiment" (Hume [1751] 1998: 85). Mackie assesses various specifications of this positive hypothesis—one of which is the objectification thesis—and it is precisely because he wants to characterize the sentimentalist hypothesis in a way that leaves open a range of theoretical options that he words (iii) so formlessly.

Mackie thinks that (i)–(iii) jointly knock out all rivals to the objectification thesis as follows (very roughly): Noncognitivist offerings fail to satisfy (i). Versions of "subjectivism" according to which moral utterances make reference *descriptively* to some real or hypothetical agent's emotions (such as the ideal observer theory) fail to satisfy (ii). Various forms of moral rationalism, intuitionism, and naturalistic realism—theories that do not accord emotion a central role in our apprehension of moral truths—fail to satisfy (iii). By contrast, the objectification thesis is supposed to pass the test. According to this thesis, moral judgment begins with humans responding to certain actions and characters in the world with affective attitudes (thus satisfying (iii)), which we then project onto our experience of the world, reading "some sort of image" of the attitude into the item that prompted it, seeing (for example) the action as categorically required (thus satisfying (ii)); and we are, by and large, fooled by this operation into thinking that the normative property really is instantiated, in which case our language for discussing it is, naturally, assertoric and propositional (thus satisfying (i)).

To assess this argument for the objectification thesis would require a comprehensive metaethical investigation. We would have to evaluate whether all these rivals do indeed fail to satisfy the desiderata that Mackie claims they fail; we would have to assess whether the objectification thesis really does satisfy the three desiderata; we would have to consider whether any theory other than the objectification thesis promises to satisfy the three desiderata; we would need to investigate whether these three desiderata really *are* that (that a theory's failure to satisfy one of these criteria really does represent a reason for rejecting the theory); and we would need to reflect on whether there might exist additional desiderata (ones, perhaps, that the objectification thesis fails to satisfy). Needless to say, such an assessment is not going to be attempted here. The point I want to observe is the general one that here Mackie is seeking to establish the objectification thesis via a metaethical route—and a long and controversial metaethical route at that. It is clear that this argument for objectification is not independent of pivotal elements of his other arguments for moral skepticism.

For example, it is a central plank of the argument from queerness that moral judgments are imbued with "objective prescriptivity"—"something that involves a call for action or for the refraining from action, and one that is absolute, not contingent on any desire or preference or policy or choice" (1977: 33). But note that this premise in the argument from queerness is really nothing more or less than desideratum (ii) employed in the argument for the objectification thesis. Many critics of the argument from queerness complain that moral discourse is committed to nothing so extravagant, and

if they are correct then the argument will clearly collapse. Whatever grounds these critics of the argument from queerness have for their view are grounds for denying that (ii) is a criterion of theory acceptance. Another class of critic of the argument from queerness will maintain that this quality of "objective prescriptivity" can indeed be satisfied by (clusters of) naturalistic properties. But saying this is nothing more or less than denying Mackie's claim that moral naturalism cannot satisfy (ii).

That Mackie's argument for the objectification thesis shares central premises with the argument from queerness is not a problem in the context of his 1980 book, for the objectification thesis is not there functioning to satisfy a proviso to another argument. But it does mean that we cannot lift the argument for the objectification thesis found in the 1980 book and use it to help establish the objectification thesis in the context of the 1977 argument. The interpretation I have offered of the 1977 dialectic has acceptance of the arguments from relativity and queerness held in abeyance until the objectivity thesis steps in as a tie-breaker, rendering it "less paradoxical to reject than to retain the common-sense belief in the objectivity of moral values" (1977: 42). But this strategy plainly will not work if whatever doubt hangs over the arguments for moral skepticism also hangs over the argument that would convince us of the thesis that would satisfy the proviso.

I conclude that Mackie's use of the objectification thesis in *Ethics: Inventing Right and Wrong* is a reasonable idea that is poorly executed. The general strategy of overcoming epistemic conservatism by showing moral intuitions to be the product of an unreliable process of formation is a sound one. Seeing moral experience as the product of an operation of projection or objectification is one prominent example of this strategy (among others) that might very well succeed. But the objectification thesis necessary and sufficient for the job is the minimal psychological thesis, to be established (if at all) by empirical investigation independent of any metaethical arguments. By muddling up psychological hypotheses with metaphysical commitments (whereby objectification involves "false belief in the fictitious features" (1980: 72)), Mackie makes it impossible for himself to use the objectification thesis in support of moral skepticism in a non-question-begging way.

10

Is Moral Projectivism Empirically Tractable?

"Projectivism" means different things to different philosophers. Even restricting our attention to *moral* projectivism merely identifies our subject matter while leaving the nature of the *projectivist* component of the thesis indeterminate. The objectives of this essay are to home in on and clarify one central thesis that seems deserving of the name "moral projectivism," and to call attention to the fact that it is an empirical hypothesis and thus must be tested as such. I should at the outset immediately quell any expectations that in this essay I will design, develop, or even suggest any experimental methods. The preliminary task of identifying and clarifying a target hypothesis is sufficiently complicated to fill the essay, and I lack the space to propose any specific empirical procedures. If what follows serves to reorient thinking about moral projectivism in an empirical direction, if it encourages people to reflect on ways by which it might be properly tested, then I will be satisfied with the contribution.

The Many Moral Projectivisms

Along with just about everyone else who discusses the topic of projectivism, I shall begin with David Hume:

> 'Tis a common observation, that the mind has a great propensity to spread itself on external objects. ([1740] 1978: 167)

"A common observation"? Restricted to the domain of philosophers of the seventeenth and eighteenth centuries, yes. A century before Hume's comment, Descartes had described humans as "accustomed . . . to attribute to bodies many things which belong only to the soul" ([1641] 1970: 109). A few years before that, Galileo declared that "many sensations which are deemed to be qualities residing in external subjects [including tastes, odors, smells, and heat] have no real existence except in ourselves, and outside of us are nothing but names" ([1623] 1960: 12). Hobbes very probably was a projectivist about many aspects of human experience, including morality (see Darwall 2000). And Newton (who was of course enormously influential upon Hume) endorsed what can be interpreted as a projectivist view of color in his best-selling *Opticks* of 1704.

In the above quote from the *Treatise*, Hume is discussing not morality but the human idea of *necessary connection*. It is generally assumed that he intends the same treatment for morals, beauty, color, sounds, and other perceptible sensory qualities. In the *Enquiry Concerning the Principles of Morals*, he apparently applies the "common observation" to morality:

> Thus the distinct boundaries and offices of reason and of taste are easily ascertained. The former conveys the knowledge of truth and falsehood: the latter gives the sentiment of beauty and deformity, vice and virtue. The one discovers objects as they really stand in nature, without addition or diminution: the other has a productive faculty, and gilding or staining all natural objects with the colours, borrowed from internal sentiment, raises in a manner a new creation. ([1751] 1998)

However, R. M. Sainsbury (1998) takes issue with the assumption that what Hume says for necessary connection is supposed to go for morality and the rest. The difference, thinks Sainsbury, is that (according to Hume) our projected idea of necessary connection leads us to massive doxastic error about the nature of reality, whereas our projected ideas of vice and virtue (and moral qualities in general) do not necessarily result in error. Sainsbury, in fact, thinks that Hume's moral projectivism is compatible with a commitment to moral realism.[1] I am not here concerned with whether Sainsbury has correctly identified a distinction present in Hume's texts; I am interested in the general distinction between error-implying and non-error-implying versions of moral projectivism. It is a distinction that should be familiar to scholars of modern metaethics. On the one hand, John Mackie—a prominent advocate of the moral error theory (and who coined the label, no less)—is a moral projectivist (1977: 42–6, 1982: 72).[2] On the other hand, Simon Blackburn—a staunch critic of Mackie's error theory and promoter of the contrary metaethical theory of noncognitivism—is also a moral projectivist (1993, 1998).[3]

So we already have three positions in play: First, moral projectivism coupled with an error theory (Mackie); second, moral projectivism coupled with realism (Sainsbury's Hume); third, moral projectivism coupled with noncognitivism (Blackburn). A conspicuous question is whether these are really three distinct variants of moral projectivism, or whether we have a single projectivism that is neutral among these metaethical options.

[1] Edward Craig (2000) also argues for the compatibility of projectivist and realist interpretations of Hume, but, unlike Sainsbury, Craig thinks that Hume can be interpreted as both regarding *causality*.

[2] Mackie tends to prefer the term "objectification." It is clear, however, that he means to capture a kind of projectivism. In his 1980 book he provides a typical description of Humean projectivism (along with the *Enquiry* quote given above), and twice refers to "this projection or objectification" (72). For discussion of Mackie's view of objectification, see Joyce (2010a).

[3] A. W. Price (1992) also distinguishes the nihilistic (error-theoretic) form of projectivism from other forms deserving the name: Simon Blackburn's "reductive projectivism" and Richard Wollheim's "genetic projectivism."

Sainsbury evidently thinks that there is more than one kind of projectivism discernible in the Humean texts. He argues that the *Treatise* relation of "spreading" (pertaining to the idea of necessary connection) is distinct from the *Enquiry* relation of "gilding or staining" (pertaining to morals, *inter alia*). The implication is that there are two kinds of projectivism: "spreading projectivism" (which entails error) and "gilding projectivism" (which does not).[4] An alternative view is that there is just one kind of Humean projectivism—that "spreading" and "gilding" and "staining" are all metaphorical synonyms—and that this projectivism and the metaphors used to describe it are simply neutral on the question of error. One might then hold that in the case of necessary connection the projectivism happens to be coupled with error, but in the case of morals it is not. The fact that Sainsbury carefully distinguishes "spreading" from "gilding or staining" indicates that this is not his view.

It may help clarify matters if we begin to break possible projectivisms down into sub-theses. (This will be done initially in a rough and ready way, later in a more rigorous manner.) Let's do it for the particular case of the idea of necessary connection.

1. We experience necessary connection as an objective feature of the world.
2. This experience has its origin in some non-perceptual faculty; in particular, upon observing a regularity in nature we form an *expectation* that brings about the experience described in 1.
3. In fact, necessary connections do not exist in the world.
4. When we utter sentences of the form "X is necessarily connected to Y" we are misdescribing the world; we are in error.

If there is a kind of projectivism that entails an error theory, then it must look something like 1–4. This, I take it, would capture Sainsbury's "spreading projectivism." I will call this "nihilistic projectivism." What of the putative non-nihilistic projectivisms?

In order to render projectivism compatible with noncognitivism, we must delete 4 (or the analog of 4). According to classic moral noncognitivism, when we utter the sentence "X is morally wrong" we are not *describing* the world at all, and therefore cannot be *mis*describing it. Blackburn, in particular, is keen to emphasize that everyday moral language is not *in error*, despite its projectivist foundation. (If there is an error anywhere, it is the blunder of philosophers who misdescribe the metaphysical commitments of moral discourse.) According to the noncognitivist advocate of projectivism, although our experience may be as of objective moral facts (as in 1), our moral language does not perform the function of expressing the belief that these moral facts obtain; rather (according to an expressivist version of noncognitivism), the function of moral utterances is to express the emotions that give rise to the experiences. (More on this later.)

But although deleting 4 results in a projectivism compatible with noncognitivism, it does not yield a projectivism compatible with realism. In order to achieve the latter, we

[4] These are not Sainsbury's labels.

need also to delete 3. This, I take it, would (*mutatis mutandis*) capture Sainsbury's "gilding projectivism"—the projectivism he associates with Hume's stance on morality.

Let us be clear. First, the conjunction of 1 and 2 is neutral between realism and anti-realism; although compatible with realism, the conjunction of 1 and 2 does not entail realism. Nor does it entail either 3 or 4. Second, the conjunction of 1, 2, and 3 entails anti-realism, but is neutral between cognitivism and noncognitivism; although compatible with noncognitivism, the conjunction of 1, 2, and 3 does not entail noncognitivism. Nor does it entail 4. Third, the conjunction of 1, 2, 3, and 4 entails anti-realism in general and an error theory in particular. Note that no combination of sub-theses *entails* realism, and no combination *entails* noncognitivism.[5]

Let us pause to think more carefully about the relation between moral projectivism and noncognitivism. First we had better alter our example to the moral case, along the same lines as we did for causal projectivism:

1. We experience moral wrongness (for example) as an objective feature of the world.
2. This experience has its origin in some non-perceptual faculty; in particular, upon observing certain actions and characters (and so on) we have an affective attitude (for example, the emotion of disapproval) that brings about the experience described in 1.
3. In fact, moral wrongness does not exist in the world.
4. When we utter sentences of the form "X is morally wrong" we are misdescribing the world; we are in error.

I have claimed that nothing here entails noncognitivism, but I can imagine someone tempted to take issue with this. "Surely," the complaint would go, "if 1 and 2 are true (and we can throw in 3 for good measure) then what lies behind moral experience is an emotion: disapproval. If one then makes public one's moral judgment—via an utterance of the sentence 'X is morally wrong'—this sentence thus expresses the disapproval. But the thesis that moral utterances express emotions just *is* noncognitivism. So in fact 4 is incompatible with 1–3; in place of 4 we should have '*Therefore*, when we utter sentences of the form "X is morally wrong" we are expressing our emotions.'"

Such an objection would be based on misunderstanding. The metaethical debate between the cognitivist and the noncognitivist does not concern what kinds of mental states *cause* moral judgments; it concerns the linguistic function of moral judgments (whether they are assertions, or commands, or interjections, and so on).[6] For S's utterance U to *express* mental state M (in the sense relevant to the metaethical debate) it is

[5] For no-frills discussion of the relation between realism, noncognitivism, and the error-theoretic stance, see my entry for "moral anti-realism" in the *Stanford Encyclopedia of Philosophy* (Joyce 2007a).

[6] Psychologists sometimes use the term "moral emotivism" to denote the theory that emotional faculties play a central role in the causal generation of moral judgment. Although philosophers are trained to think of emotivism as a kind of noncognitivism, it is clear that in this case the taxonomy does not apply; psychologists are not using "emotivism" in its metaethical sense. See Joyce (2008) for analysis.

neither necessary nor sufficient that U is caused by M. What concerns us here is the sufficiency condition. From the fact that an emotion is causally active in the generation of moral judgment it does not follow that the moral judgment *expresses* that emotion—not, at least, in the relevant sense of "express." Suppose my child desperately wants a bike for his birthday, and I am inclined to buy him one because I love him and want him to be happy. This love is a central component in the causal chain leading up to me uttering the sentence "I intend to buy you a bike." And yet this utterance is a straightforward assertion: It has truth conditions and functions linguistically to express a belief (the belief that I intend to buy him a bike). The fact that an emotion has caused the utterance should not distract us from being all-out cognitivists about this utterance. Thus, although sub-theses 1 and 2 add up to the claim that affective attitudes (for example, disapproval) are causally generative in moral judgment—perhaps even necessary for moral judgment—1 and 2 (even with 3 thrown in for good measure) fall short of entailing noncognitivism. 1–3 together *are silent* on the function of public moral utterances.[7]

There is, nevertheless, a tendency in some quarters to think that projectivism entails noncognitivism. Perhaps this is based on the misunderstanding just diagnosed. Or perhaps there is yet another kind of projectivism possible—one that really does entail noncognitivism. Consider these statements by Nick Zangwill, both describing Blackburn's view:

> "Projectivism" is the view that the disputed judgements express non-cognitive mental states, such as emotions, desires, habits, or expectations; but the projectivist also holds that such non-cognitive states are *spread* or *projected* onto the genuine facts and states of affairs. So we come to speak and think as if there were an extra layer of properties in the world. (1992: 161)

> According to Simon Blackburn . . . "projectivism" . . . is the view that moral judgements express attitudes (approval, disapproval, liking or disliking, for example), which we "project" or "spread" onto the world. (1990: 583)

Note that both characterizations render projectivism a double-barreled thesis. The second component seems familiar: One could plausibly see the idea of "non-cognitive

[7] The necessity condition also fails. A speech act (such as an assertion, or an apology, or a promise, or an interjection, or a command) may express a type of mental state (such as a belief, or regret, or a commitment) without the speaker having that mental state. One need reflect only on the phenomenon of *insincerity* to see this. An insincere act of promising still succeeds in being a promise (unlike, say, an overtly sarcastic promise utterance, which is not a promise at all), and, as such, an insincere promise still expresses a commitment on the speaker's behalf—a commitment that the speaker actually does not have at the time of utterance. Similarly, an insincere assertion (a lie) expresses a belief that the speaker does not have at the time of utterance. But if speaker S can, through uttering U, express mental state M *while S does not have M* (and perhaps has not had M), then the relation between M and U cannot be a causal one. Rather, the relation is a complex one concerning the entrenched linguistic conventions understood by both speaker and audience. If both speaker and audience take it that acts of promising express commitment—if this understanding is a necessary prerequisite to being granted competence with the concept *promising*—then a successful act of promising does express commitment, irrespective of whether the speaker as a matter of fact is committed. (I seem to have said this many times before; in the unlikely event that anyone has been paying attention, I apologize for the repetition.)

states . . . spread or projected onto genuine facts and states of affairs" as a pithy summary of 1 and 2. But the first component is one that we haven't yet encountered as a sub-thesis of moral projectivism. This first component explicitly packs the noncognitivist case into the definition of projectivism: "disputed judgments express non-cognitive states, such as emotions, desires [and so on]." This, of course, excludes the possibility of projectivism being compatible with either realism or the error theory.

We now potentially have four kinds of moral projectivism to deal with, to which we can, for the sake of convenience, give the following labels:

- Minimal projectivism: the conjunction of 1 and 2.
- Metaphysical projectivism: the conjunction of 1, 2, and 3. (I call this "metaphysical" because the addition of sub-thesis 3 adds a metaphysical claim that otherwise isn't present.)
- Nihilistic projectivism: the conjunction of 1, 2, 3, and 4.
- Noncognitivist projectivism: the "double-barreled" thesis just described, which is minimal projectivism (or possibly metaphysical projectivism) conjoined with expressivist noncognitivism.

	Thesis 1	**Thesis 2**	**Thesis 3**	**Thesis 4**	**Expressivism**
Minimal projectivism	√	√	–	–	–
Metaphysical projectivism	√	√	√	–	–
Nihilistic projectivism	√	√	√	√	x
Noncognitivist projectivism	√	√	?	x	√

[NB: The last two columns exclude each other, hence the "x" is to be interpreted as a *denial* (as opposed to simply the absence of endorsement).]

One form of projectivism entails noncognitivism, one excludes noncognitivism, and two are silent on the matter. I do not propose to adjudicate among these theoretical options; it is possible that different kinds of projectivism—even different kinds of *moral* projectivism—are better suited than others for different theoretic purposes. However, I would like to express my misgivings about the noncognitivist projectivism described by Zangwill, for it seems to me to contain an inherent tension. Start by focusing on the second barrel of the first passage: "we come to speak and think as if there were an extra layer of properties in the world"—a comment that jibes with remarks made by Descartes, Galileo, Hume, and the rest. But what is it to "speak and think" as if the world were a certain way? The natural answer is that speaking as if the world were a certain way amounts to *asserting* that it is that way, and thinking as if the world were a certain way amounts to *believing* that it is that way. But asserting and believing are the hallmarks of a *cognitivist* attitude. In other words, the second barrel of noncognitivist projectivism (as stated by Zangwill) seems to presuppose the misfiring of the first barrel. Granted, the tension here is not flat out inconsistency. It is possible to "speak and think" as if the world were a certain way without asserting and believing

that it is. A paradigm example of this phenomenon would be acting. The actor utters (and, presumably, thinks) the sentence "Thou art a scholar" without asserting or believing the proposition. So it is possible to reconcile the first and second barrels of this form of moral projectivism, but one will have to tell a special story about speaking-without-asserting and thinking-without-believing in order to do so.

A similar issue arises when we give consideration to whether minimal projectivism entails cognitivism. (Indeed, to the extent that the second barrel of Zangwill's noncognitivist projectivism is a rough statement of minimal projectivism—or possibly metaphysical projectivism—it is the very same issue.) One might be tempted to assume that sub-thesis 1 alone entails cognitivism, for 1 consists of a claim about the nature of moral experience, which, one might think, implies something about the subject's *beliefs*. Doesn't sub-thesis 1 amount to the claim that we *believe* moral wrongness to be an objective feature of the world? No, it doesn't. Let us distinguish between moral experience and moral judgment. There is a clear sense in which a stick in water is experienced as bent, but the savvy observer does not judge it (believe it) to be bent. One can imagine a world where minimal projectivism is true but subjects are more or less aware of the fact and are not fooled. Perhaps the inhabitants of this world are all moral error theorists but they continue to have experiences as of an objective moral realm, which they treat as a kind of unavoidable mirage. Or perhaps the inhabitants of this world have located some facts of a subjective (for example, constructivist?) nature, which they judge worthy referents of their moral vocabulary. These people find that they cannot help but experience these facts as objective, but they know that are not; they do not treat the moral realm per se as illusory, but they treat the objective pretensions of that realm as an ineluctable illusion. So sub-thesis 1—even coupled with 2 to form minimal projectivism—does not entail cognitivism. Nevertheless, it seems fair to say that *generally speaking* how people think and speak about the world goes along with their experience of the world. If people are experiencing the world as containing objective moral properties—as sub-thesis 1 declares—then it is natural to suspect that people will *believe* the world to contain such properties, and that their language will reflect these beliefs by consisting of *assertions* that the world instantiates such properties. As before, the burden seems to fall on the noncognitivist to tell a special story about how ordinary persons have a critical distance from their moral experience—how they possess a certain kind of sophistication—such that they do not take their experience at face value.

In other words, it seems as if moral cognitivism is the "natural partner" of minimal projectivism—the default assumption—and that if one wants to endorse minimal projectivism while supporting moral noncognitivism, then one has some explaining to do. The important thing to note is that the projectivist cannot have it both ways: She cannot maintain both (a) that our moral experience, our tendency to "spread" our emotions onto reality, really *fools* us, and (b) that in making a moral judgment we are doing no more than expressing our emotions. What it is for our experience to "fool us" is, presumably, for our *beliefs* to go along with how things seem (in which case *assertion* comes naturally along too); but if we are in the business of believing (and asserting),

then it is not the case that we are doing no more than expressing our emotions (even if it is true that emotional episodes play a central causal role in the production of our moral experience); we are, rather, squarely in the province of cognitivism.

Getting straight on what kind of moral projectivism is under discussion on any given occasion is of the utmost importance, for it is a truism that before we can embark on investigating whether projectivism is true, we need first decide the content of the thesis whose truth we are scrutinizing. Different data will bear on whether different versions of projectivism are true. In order to investigate the truth of noncognitivist projectivism, for example, we would have to engage with the debate over cognitivism versus noncognitivism. But this debate can be ignored if we are investigating other kinds of projectivism. Alternatively, in order to confirm or disconfirm metaphysical projectivism we would need to establish to our satisfaction either that there are or are not objective moral facts; whereas if our interest is in minimal projectivism this does not matter.

In what follows I want to focus on sub-theses 1 and 2, which together comprise what I have labeled "minimal projectivism." Despite my calling it this, I am officially agnostic as to whether the conjunction of these two theses really ought to be categorized as a type of "projectivism" or whether it should, rather, be thought of as the common heart of all other kinds of moral projectivism. Perhaps instead of being presented as a delineation of different kinds of moral projectivism, the preceding discussion would be better interpreted as an exposé of the *confusion* surrounding the idea. Perhaps instead of living with lots of different kinds of moral projectivism, we would do better if we decided on a single unified theory. Whether that all-purpose moral projectivism would be equivalent to what I have labeled "minimal projectivism," or whether those two sub-theses should ultimately be categorized as necessary but not sufficient conditions for moral projectivism, is not something I aim to decide here. There are not, after all, any *facts* about what "projectivism" denotes to which we can appeal to settle these questions; the matter is a pragmatic one, concerning which notion of moral projectivism will serve our theoretic purposes most usefully. This noted, I will, if only for the sake of brevity, continue to refer to this conjunction of sub-theses as "minimal projectivism."

Turning a Philosophical Metaphor into an Empirical Hypothesis

It is remarkable how rarely, in all the discussions of moral projectivism over the years, proper attention has been given to the fact that the theory is generally presented and thought about in *metaphorical* terms (such as "gilding or staining").[8] Even the appellation

[8] A notable exception is D'Arms and Jacobson (2006). Simon Blackburn is also no doubt aware of the metaphorical status of references to "projection," yet (in my opinion) he has done little to replace the metaphor with a precise literal hypothesis. On at least one occasion he confesses that "projectivism" is not an entirely happy term for the position he has so frequently advocated (Blackburn 1995: 36).

"projectivism" is metaphorical, for nobody thinks that when a person projects her anger onto her experience of events (say), this emotion literally flies forth from her brain and laminates the world. (Slogan: Projecting emotions is not like projectile vomit.) But how do we determine whether a metaphor is adequate, especially since (it is usually accepted) all metaphors are by definition false? Evidently, projectivism is a theory in need of translation into literal terms before it can be properly assessed. Drawing attention to the conjunction of sub-theses 1 and 2 is an attempt to accomplish this.

Another striking feature of moral projectivism that has never, to my knowledge, been properly appreciated is that, to the extent that we can detect something literal lying behind the traditional metaphors, it seems reasonably clear that we are dealing with a thesis that is, either entirely or in part, *empirical*. Once we get past the metaphorical level, we see that projectivism concerns a claim about the nature of moral experience (sub-thesis 1)—which appears to be an empirical matter—and a claim about the genealogy of that experience (sub-thesis 2)—which also appears to be an empirical matter. Perhaps, as I say, we should decide that there is more to moral projectivism than just these two claims—so perhaps projectivism will turn out to be not *entirely* an empirical matter—but at the very least it is significant to recognize that a substantive empirical inquiry is an important necessary component of any serious attempt to assess the truth of moral projectivism.

Let us discuss these two sub-theses in turn. Doing so requires that we come up with better labels. I will call sub-thesis 1 "the phenomenological thesis" and sub-thesis 2 "the causal thesis."

The Phenomenological Thesis

1. We experience moral wrongness (for example) as an objective feature of the world.

Many metaethicists accept the phenomenological thesis. In the debate between the moral realist and her opponents, it is often taken for granted by both sides that the moral anti-realist faces a burden of proof, inasmuch as it *seems* to us that moral judgments track objective qualities. Moral realists often argue that this represents some kind of burden of proof that the anti-realist must overcome; they argue for moral realism on the basis of the combination of the phenomenological thesis with a methodological principle of epistemic conservatism. For example, we read this from Jonathan Dancy:

> [W]e take moral value to be part of the fabric of the world; taking our experience at face value, we judge it to be the experience of the moral properties of actions and agents in the world. And if we are to work with the presumption that the world is the way our experience represents it to us as being, we should take it in the absence of contrary considerations that actions and agents do have the sorts of moral properties we experience in them. This is an argument about the nature of moral experience, which moves from that nature to the probable nature of the world. (1986: 172)

And this from David Brink:

> We begin as (tacit) cognitivists and realists about ethics… We are *led* to some form of antirealism (if we are) only because we come to regard the moral realist's commitments as untenable, say, because of the apparently occult nature of moral facts or because of the apparent lack of a well developed methodology in ethics… Moral Realism should be our metaethical starting point, and we should give it up only if it does involve unacceptable metaphysical and epistemological commitments. (1989: 23–4)

The soundness of this burden-of-proof argument is not relevant to our present purposes; I am just noting the endorsement of the phenomenological thesis inherent in this popular form of argument.[9] (For the most developed version of this argument for moral realism, see Huemer 2005.[10])

The phenomenological thesis is also employed as a premise in arguments favoring moral anti-realism. John Mackie argued that not only is our moral experience as of objective values, but that this objectivism is embedded "in the meanings of moral terms" (1977: 31)—that the assumption that moral values exist objectively "has been incorporated in the basic, conventional, meanings of moral terms" (1977: 35). Mackie goes on to argue that this aspect of morality is in fact not satisfied by the world, and hence he advocates moral skepticism. Again, it is not my intention to evaluate this argument, but rather to note the central role that the phenomenological thesis plays in it.

Nevertheless, for all the widespread support enjoyed by the phenomenological thesis, it has never been properly subject to empirical scrutiny.[11] Perhaps the reason for this is that the thesis is seriously unclear on several dimensions, all of which would need to be settled before testing could be undertaken. (I doubt, however, that this typically *is* the reason, since many moral philosophers seem happy to endorse the thesis without worrying about, or attempting to settle—or even, apparently, noticing—the lack of clarity.) There are three conspicuous places where the phenomenological thesis needs refining. First, what is it to experience morality as *objective?* Second, what is it to *experience* morality as objective? Third, what is it to experience *morality* as objective? I do not propose to attempt to settle these questions here, but rather to identify what would need to be settled before anybody should pass judgment one way or the other on the phenomenological thesis. What follows are but preliminary notes.

[9] For skepticism about the burden-of-proof argument, see Kirchin (2003); Loeb (2007).

[10] I offer some criticisms of Huemer's view in Joyce (2009c).

[11] Those who have come closest are Nichols and Folds-Bennett (2003) and Goodwin and Darley (2007). One might also reasonably claim that the extensive empirical research program concerning the moral/conventional distinction (in developmental psychology, clinical psychology, and cross-cultural studies) has bearing on the phenomenological thesis, to the extent that judgments concerning moral transgressions are taken to be those that (*inter alia*) hold irrespective of any authoritative decree, which is one way of understanding objectivity. (A reasonable starting point for this large literature is Nucci 2001, Smetana 1993, and Turiel et al. 1987.) However, even these interesting studies do not target the hypothesis that we *experience* morality as objective (as opposed to believe that it is). In my opinion, Goodwin and Darley also employ a misguided notion of *objectivity*.

Objectivity

First, we must become clearer on what kind of *objectivity* is relevant to the phenomenological thesis, for this term is used in different ways by different philosophers. Michael Smith uses "objectivity" to refer to the possibility that moral questions have a correct answer upon which open-minded and clear-thinking agents will converge (1994: 5–6). Crispin Wright associates objectivity with whatever plays a wide (as opposed to narrow) cosmological role (1992).[12] Michael Dummett, by contrast, prefers to argue that sentences of a certain kind are objective if and only if we think of them as determinately true or false, though we nevertheless know of no method representing either a proof or a disproof (that is, the sentences are potentially "recognition transcendent") (1978, 1993). "Objectivity" is often associated with some notion of *mind-independence*, though the matter is far from straightforward, since there are numerous kinds of mind-(in)dependence relations possible. (Cars, for example, are generally classified as concrete, mind-independent entities, despite the fact that they were designed and built by and for creatures with minds.) We can contrast *existential* mind-independence (X would exist even if no minds existed) with *conceptual* mind-independence (the concept *X* can be adequately articulated without making reference to any mental entities). For example, if one were to hold that the correct analysis of the concept *moral goodness* is something of the form "*whatever an observer with qualities Q would approve of in circumstances C*," this would make moral goodness existentially mind-independent (since its instantiation would not depend on the existence of any such observers) but conceptually mind-dependent (since *approval* is a psychological category ineliminable from the explication). In the case of morality, there is the possibility of a further kind of *practical* objectivity: namely, that moral imperatives have a distinct kind of categorical authority: Maurice Mandelbaum writes that our feeling of being bound by a moral obligation "appears as being independent of preference . . . as an 'objective' demand" (Mandelbaum 1955: 50).

One response to this abundance of non-equivalent notions of *objectivity* is to judge that what is called for is further discrimination of different varieties of moral projectivism, depending on which distinct kind of objectivity is built into the phenomenological thesis (and consequently into the causal thesis). But I think this would be a profligate and implausible response. It seems unlikely, on the face of it, that just anything that has been given the moniker "objectivity" by philosophers—however legitimately for their local purposes—can be plugged into the phenomenological thesis while still yielding a recognizably *projectivist* theory. It is more plausible that the intuitions lying behind projectivism will be best captured by homing in on a particular kind of objectivity, or a cluster of related kinds of objectivity. It is even possible that upon further reflection we

[12] A subject matter has a wide cosmological role if the kinds of things with which it deals figure in a variety of explanatory contexts—specifically, if they explain things other than (or other than via) our judgments concerning them.

may prefer to eliminate the word "objectivity" altogether in favor of something more unequivocal.

What seems common to all brands of projectivism is that something-or-other is experienced as "out there," existing, or having certain qualities, antecedently and independently of the subject. Let us postpone the question of how we can *experience* something as "out there," and just focus on the "out-there-ness" itself. We tend to think of cats, rocks, tables, planets, relative size, chemical constitution, and duration as features of the world, independently of our act of perceiving them. Even if a person has directly *caused* a particular cat's existence (via arranging a breeding program, say), there is still a robust sense to be attached to the idea that the person does not "constitute" the cat's existence in the act of apprehending it. But is there anything that we *don't* think of in this way? Gideon Rosen (1994) has argued that there is little sense to be made of this dichotomy of objectivity/subjectivity—at least with respect to the ubiquitous role it has traditionally played in philosophical debates.

> To be sure, we do have "intuitions" of a sort about when the rhetoric of objectivity is appropriate and when it isn't. But these intuitions are fragile, and every effort I know to find the principle that underlies them collapses. We sense that there is a heady metaphysical thesis at stake in these debates over realism . . . [b]ut after a point, when every attempt to say just what the issue is has come up empty, we have no real choice but to conclude that despite all the wonderful, suggestive imagery, there is ultimately nothing in the neighborhood to discuss. (1994: 279)

Consider the emotion of sadness, which is clearly mind-dependent in a perfectly trivial sense.[13] And yet for any given agent there are many instances of sadness (nearly all of them) of which she is not the author: Nearly all episodes of sadness, for any person, are items "out there," subjects of discovery, not things she invents or creates in the act of perceiving them. Thus Rosen would doubt that even in the case of sadness have we succeeded in "abrogating the right to think of these facts as robustly real constituents of a mind-independent order" (293). He describes several attempts to frame the distinction—several ways of understanding what it might mean for something to have a "less-than-objective" ontological status—and each he rejects because of the persistent availability of the "anthropological perspective": For any putatively "subjective" phenomenon, Rosen will imagine an anthropologist investigating it, and he will observe that even though the phenomenon may ultimately supervene on psychological states (for example, pretty much any phenomenon that is the topic of any of the social sciences), the anthropologist nevertheless is, from her own perspective, engaged in the study of a robustly real part of the natural world order.

My purposes here do not require that Rosen's arguments be countered (though I have briefly critically discussed them in Joyce 2007a, and see note 15 below); it suffices to note that what we are seeking in trying to clarify the phenomenological thesis of projectivism is something considerably more modest than defending the broad

[13] By restricting attention to "the *emotion* of sadness," I hope to put aside tricky (but clearly different) cases involving sad music, sad events, sad faces, and so on.

imagery of *objectivity versus subjectivity* that motivates so many philosophical debates. To make clear that we are stipulating a notion just for our local purposes, let me use the term "subject-(in)dependence," rather than "mind-(in)dependence." The important thing to note is that we can define *subject-(in)dependence* to be a relativistic notion. Consider a particular episode of sadness: say, Sally's sadness on Tuesday afternoon. This sadness is subject-dependent relative to Sally; it is subject-independent relative to everyone else. Even for Sally's boyfriend, who (let us assume) *caused* the sadness, Sally's sadness is an item in the world (albeit a psychological item) that is there to be discovered, of which he might be ignorant, of which he is a passive observer, which *could* have occurred without him. There are important questions to answer still—most prominently, what relation precisely does Sally bear to this episode of sadness in virtue of which it is subject-dependent, relative to her?[14]—but I do not propose to pursue them here. I would rather point out the virtues of taking the path of relativism on this matter. First, by making subject-(in)dependence a relativistic notion we have nullified the significance of the anthropological perspective. If a phenomenon is subject-independent relative to an investigating anthropologist, so what?—it may nevertheless be pertinent to note that (unlike many phenomena) it is subject-dependent relative to some other individual.[15] Second, and more importantly, employing a relativistic notion is all we need. After all, what we intuitively want to capture of the projectivist tendency is the experience a person may have that "I am not the author of this phenomenon; it would carry on the same even unperceived by me." We do not have to worry about the absolute "objective" status of the phenomenon (or, indeed, whether it is even coherent to think of any such notion of absolute objective status standing in contrast to absolute subjective status); we need concern ourselves only with how the subject experiences it *in relation to herself*.

[14] I am more comfortable saying something about what this relation *does not* consist in. It does not consist in Sally *causing* the sadness, and it does not consist in Sally judging or believing that she is sad. (I am willing to accept that she may be sad without believing herself to be.) I am tempted to cash it out in terms of *a priori* modal dependence. *This* token episode of sadness (had by Sally on Tuesday afternoon) *could only* have been had *by Sally*. If we imagine a possible world, W, strikingly similar to ours—where there is someone very much like Sally, feeling sadness in very similar circumstances (on Tuesday afternoon, and so on)—but for which we stipulate that (for whatever minimal reason) she is in fact *not* Sally (and does not count even as her modal counterpart, despite the similarities), and nor is anyone else at W, then we would (I suggest) conclude *a priori* that *this* token episode of sadness (gesturing to the actual Sally's actual sadness) does not exist at W. I confess, though, that I am not at all confident that this thinking will produce the intuitively correct output across all cases we might want to consider.

[15] Incidentally, from this relativistic notion we could then build an absolute one: A phenomenon is Subject-Independent in the absolute sense (note the upper case) iff there is no perspective relative to which it is subject-dependent. This seems to be something Rosen overlooks. It may well be that for any "subjective" phenomenon we can invoke the anthropological perspective (thus, he thinks, casting the objective/subjective distinction into disarray), but the reverse does not hold. It is not the case that for any "objective" phenomenon (say, the chemical constitution of Jupiter) we can with equal ease invoke the "subjective" perspective, from which some mental activity constitutes the facts of the case. For all Rosen's arguments, we can still distinguish those cases for which discovery-talk and mind-dependence-talk can co-exist from those cases for which discovery-talk is permissible but mind-dependence-talk is wholly misplaced—and this distinction may be of philosophical significance.

A third point to reflect upon is the possibility that the kind of "out-there-ness" that the folk employ in their judgments of objectivity is inchoate and in fact resistant to more precise analysis. Although it is natural for a philosopher to seek a more precise understanding of what it is to for a phenomenon to be subject-(in)dependent (relative to an agent), if the everyday notion that figures in people's thinking is in fact indeterminate, then (a) for the purposes of gauging whether a token judgment is imbued with "out-there-ness" it may not be *necessary* to precisify the relevant notion of subject-independent "out-there-ness" beyond a nebulous and coarse-grained version, (b) it may not be *desirable* to so precisify the notion (since we want to ensure that we are capturing the folk idea), and (c) it may not even be *possible* to precisify the notion more than we have. Regarding (c), it may turn out that Rosen is entirely correct about the ineliminable confusions lying at the heart of the objective/subjective distinction, but this would not show that people do not *employ* the notion (warts and all). Let us not forget that our task is not to produce a philosophically defensible characterization of *objectivity*—not even as it appears in the phenomenological thesis. Our task is to identify what notion the folk are utilizing, sufficient for us to distinguish those circumstances where they employ it from those circumstances where they do not. If in fact the folk are deeply confused, then a deeply confused notion is the one we should be isolating.[16]

Experience

There is much that remains to be clarified in what has just been said, but already the second disambiguation of the phenomenological thesis cries out for discussion. How, it might be asked, can this kind of "out-there-ness" possibly be the object of *experience?* Surely (the objection goes) what we experience is far more primitive and simple than anything remotely like this? It is, however, highly debatable how meager or rich the content of experience is. Even confining ourselves to *visual perception*, it has been argued that the contents of experience can include relatively thick properties, such as *being caused by, being an object, being a house* (*a tree*, and so on), and *being subject-independent* (see Searle 1983; Siegel 2006a, 2006b). The last is of particular interest here. One view is that visual experience incorporates *expectations* of how something may change relative to the viewer, and these expectations constitute a phenomenality of subject-(in)dependence; the counterfactuals are not merely beliefs formed by the subject on the basis of visual data, but are properly thought of as part of the visual experience itself (see Merleau-Ponty 1945; O'Regan and Noë 2001). It would be a project of much interest to see whether this line of thought could be plausibly extended to moral experience.

[16] Those with reservations that the folk could possibly be employing a deeply confused or inchoate and indeterminate notion might recall how Socrates typically sets out to demonstrate exactly this: that despite confidently employing a term like "justice" or "knowledge," his interlocutors in fact do not really have any precise idea what they are talking about.

However, it is not necessary to our present purposes to undertake anything so controversial; we do not need to support the view that subject-independence can be experienced *by the senses*. It suffices if subject-independence can figure in mental states—whether these states be perceptual, perceptual based, or otherwise—and there is something *that it is like* to have these mental states. Suppose that the mental states in question are just common-or-garden *beliefs;* there is still a strong case to be made that there is something *that it is like* to have such beliefs occurrently. (See Flanagan 1992; Goldman 1993; Peacocke 1999; Horgan and Tienson 2002; Kriegel 2003; Pitt 2004.)[17] (This option may be unavailable to the noncognitivist projectivist, who may be inclined to doubt the existence of moral beliefs entirely.)

It is also worth noting that the term "experience" is often used in more liberal ways that may have little to do with any state for which there is a fact about what it is like to have it. Consider: "She experienced the fall of Paris in 1940," "He experienced the bad weather as a personal slight," "He experienced his mother as overbearing and critical," "She experienced a great deal of opposition to her project," "The stock market experienced a slump," "New Orleans experienced heavy rainfall." Clearly, there is much variation among these uses; I shall not attempt to classify or analyze them. The point is that there is enough pliability to the term "experience" that we should not too quickly assume that we know what its appearance in the phenomenological thesis amounts to. Specifically, it may be a gloss for something like "We have strong intuitions that moral wrongness (for example) is an objective feature of the world." Perhaps such intuitions have some kind of "what-it-is-likeness" to them; *perhaps they do not*. But even if lacking phenomenal character, sufficiently spontaneous and entrenched intuitions may serve to underwrite the first sub-thesis of moral projectivism.

Consider, for example, the notion of projection that is often employed in psychopathology, such as when a subject is said to project his feelings as a defense mechanism. Perhaps the example from above, of a person experiencing his mother as critical and overbearing, would be an apt illustration to use. Let us say that his mother is not actually critical and overbearing at all (sub-thesis 3 of projectivism), but that the subject's feelings in this respect are caused by his own sense of inadequacy (sub-thesis 2). Now, although I have just used the words "experience" and "feelings" with regard to the subject's attitude toward his mother's (supposed) personality, does the projectivist case here really depend on these states having a "what-it-is-likeness" to them? I shouldn't think so. Perhaps the subject simply *believes* his mother to be like this, and perhaps (contra the views mentioned above) these beliefs have no quality that can be legitimately called "phenomenal character." This in itself would not undermine the psychopathological diagnosis that the subject is projecting his sense of inadequacy in his dealings with his mother. Similarly, it has been shown that persons suffering from

[17] For what it is worth, Hume did not think that anything like "subject-independence" was part of the content of sensory experience: "[A]s to the *independency* of our perceptions on ourselves, this can never be an object of the senses" ([1740] 1978: 191).

forms of social phobia have a variety of distorted beliefs about social interactions, including assigning a high probability to the proposition that some social gaffe will be committed (Newmark et al. 1973). It seems plausible to say that such persons are projecting their fears onto their view of possible events. Yet the plausibility of this claim does not hinge on the mental act of assigning a high probability to certain events having a phenomenal character.[18]

In light of these last comments, perhaps I have named the phenomenological thesis poorly. Yet my use of the term reflects a liberal attitude in metaethics in general (or so it seems to me): Philosophers often refer to "moral phenomenology" meaning "how morality *seems*," without discussing or even assuming that this "seems" has any phenomenal character in the sense that philosophers of mind intend the phrase. We all know that there are uses of "seems" that do not presuppose phenomenality—for example, "It seems that dinosaurs went extinct 60 million years ago" (see Tolhurst 1998). I do not, for example, think that the quotes by Dancy and Brink given earlier indicate an intention on their parts to commit to a strong and literal sense of moral phenomenal character.

Even supposing that we *are* talking about the phenomenal character of moral judgments, it is important to note that we are not presupposing that there is something *distinct* about moral phenomenality. It has been noted that moral phenomenology is an approach that presupposes that there is something *peculiar* about the phenomenal quality of morality, such that if there is not, the whole pursuit becomes spurious (see Sinnott-Armstrong 2008; Kriegel 2008). Investigating the phenomenological thesis of projectivism does not engage us in that approach. We are interested in whether moral judgments have a quality of subject-independence. Perhaps judgments about many other things have this quality too (judgments about cats, rocks, tables, planets, relative size, chemical constitution, and duration). Perhaps there is nothing special about the kind of subject-independence that is attributed to morality (assuming that it is). Indeed, one obvious method for investigating whether moral judgments are imbued with subject-independence is to look for similarities—perhaps even exact matches—with other kinds of subject-independent judgment. The phenomenological thesis asserts a simple

[18] I am not claiming that such non-phenomenal "experience" *must* take the form of belief. I should like to maintain the earlier distinction between experience and judgment: One can experience something *as X* while judging that it is *not X*. The introduction of a more fine-grained framework that would accommodate this does not seem objectionable. Let us further consider the phobic, though we will change the example to an arachnophobe. Suppose therapy leads the arachnophobe to understand his problem; he comes to realize (all things considered) that the spiders he encounters pose no threat. (We shall assume that he does not live in Australia!) Yet, when he comes upon a daddy long-legs in the bathtub, he finds himself once more in the grip of the thought that the spider is (in some possibly inchoate sense) dangerous. We might choose to accord this "thought" some phenomenal quality (and of course for the phobic this thought is also accompanied by anxiety, which surely does have a phenomenal flavor to it), but doing so does not seem compulsory. Even so, it seems desirable to distinguish the phobic's thought from a straightforward *belief*. Arguably, the phobic has ceased genuinely to *believe* that the spider is dangerous; he just cannot help entertaining the thought. (For some discussion of the role of thoughts and beliefs in phobias, see Joyce 2000.) Maintaining some logical space between non-phenomenal "experience" and belief also helps to make this way of explicating the phenomenological thesis available to the noncognitivist projectivist (who generally denies the existence of moral *beliefs*).

predication. To think that it implies that there is some special phenomenological "signature" of morality is like thinking that someone who declares that crocodiles are green is committed to there being some distinctive greenness peculiar to crocodiles alone.

Morality

Let me now turn to the third disambiguation of the phenomenological thesis: What does it mean to say that we experience *morality* in such-and-such a manner? There are two kinds of clarification that one would ideally like to see made. First, there is disagreement over how we are to demarcate the moral from the non-moral realm. Is moral normativity necessarily distinct from prudential normativity (as Kant thought)? Are moral imperatives those that one is willing to universalize (as R. M. Hare thought)? Must moral norms concern interpersonal relations (as Kurt Baier thought)? Philosophers argue about such things, and to the extent that these disputes remain unsettled, so too does the domain of the moral. But even if we were to suppose that such worries could be resolved, we would face a second kind of indeterminacy about morality: its tremendous variation. We must distinguish moral decisions from moral judgments (Sinnott-Armstrong 2008); moral judgments of value from moral judgments of duty; moral judgments applied to oneself from moral judgments about others; direct moral judgments from removed moral judgments (Mandelbaum 1955: chapters 2 and 3); first-order from higher-order moral judgments (Horgan and Timmons 2008); moral judgments involving thin evaluative concepts (*good, bad, right, wrong*, and so on) from moral judgments involving thick evaluative concepts (*heroic, sleazebag, wimpy, fair, humiliating*, and so on); and so forth.

I have attempted to address the first kind of problem elsewhere (see, especially, Joyce 2006a, chapter 2), so will not rehearse that thinking again here. Even without settling such disputes, however, it might suffice for our present purposes if we observe that pretty much all parties will agree to certain *paradigms* of moral judgments. When an ordinary person responds to a documentary on Nazi war crimes with the utterance "Those *evil* bastards!" we will all agree that a moral judgment has occurred, even if we disagree on what qualities were present that warranted this verdict.[19] (We will also agree on *foils* of moral judgments: Nobody is arguing that someone who utters "Taking the bishop with your rook is the best move" has made a moral judgment.) Having a substantial number of paradigms in hand should suffice to test the phenomenological thesis. Finding that the phenomenological thesis holds true of all such paradigms would not, of course, allow us to conclude that it holds of *all* moral judgments, but it would at least be a substantial and interesting start.

We come to a similar conclusion when giving consideration to the second problem. If we were to discover that the phenomenological thesis holds true of, say, direct first-order

[19] I say "ordinary person" to exclude certain philosophers, who may hold all sorts of wacky views. Consider what Bishop Berkeley took himself to be saying when he uttered "There is a tree in the quad."

judgments about one's own moral duties, we should certainly not conclude that it will also hold true of second-order judgments about moral values. It is entirely conceivable that some but not all of the items on this inventory of moral types will satisfy the phenomenological thesis. But does the phenomenological thesis (or any relevant disambiguation of the phenomenological thesis) really purport to embody a claim about *all* moral experiences? A couple of paragraphs ago I dismissed the supposition that in investigating the phenomenological thesis we are seeking something *distinct* about moral experience; now the question is whether we are seeking something *common* to all moral experience. It is natural to suppose we might be; it is natural to read the phenomenological thesis as a universal generalization. This is in fact something I intend to resist, but I will postpone the matter until after I have discussed the causal thesis.

A more comprehensive essay would now present ideas on how the phenomenological thesis should be tested. That, however, is not my purpose on this occasion. I am satisfied to call attention to the fact that the thesis does amount to an empirical claim, and the only reason one may have for assuming that it cannot (in principle) be tested using scientific methods is thinking that there is something vague, vacuous, ambiguous, or incoherent about the thesis. I hope that the preceding comments have gone some way to answering those harboring any of the latter worries: I have tried to show how the thesis can be disambiguated, and I would be surprised if anyone were to think that, so clarified, something incoherent remains buried in the thesis.[20] To those persuaded that we now have an empirically testable hypothesis under consideration but who ask "OK, but *how*?" I reply "Good question; let us try to think of a good answer."

It is possible that an adequate investigation of the phenomenological thesis will at some point involve an examination of something deserving the name "intuitions"—and if this is so it is vital to avoid the pitfall of assuming that one's own intuitions—honed by years of metaethics and dripping with theoretic prejudice—should stand in for those of everyone else. If we do attempt to collect a sample of others' intuitions, it is equally vital that we do not do so in a shoddy manner (for example, questionnaires to one's *Ethics 101* undergraduates at the end of class). Data must be collected in an intellectually responsible manner, complying with the customary scientific standards concerning such things as sample size, control groups, replicability, randomization, correcting for framing effects, and so on.

The Causal Thesis

2. This experience—of morality as an objective feature of the world—has its origin in some non-perceptual faculty; in particular, upon observing certain actions

[20] Perhaps certain forms of *objectivity* have been written off as incoherent (see, for example, Rosen 1994), but, as was noted earlier, the phenomenological thesis does not require that any general concept of *objectivity* ultimately makes sense.

and characters (and so on) we have an *affective attitude* (the emotion of disapproval, for example) that brings about the experience described in 1.

Investigating the phenomenological thesis promises to be very challenging. But on the assumptions that it can be done satisfactorily (which is, of course, a big assumption), and that the hypothesis receives confirmation, then testing the causal thesis promises to be somewhat more straightforward. Let us suppose that our experience with exploring the phenomenological thesis has left us with a test, or series of tests, that we can apply to subjects in order to gauge their score on an "objectivity scale" with various kinds of experience. (This is almost certainly an idealization of anything we can reasonably hope for, but let us allow ourselves to speak in idealized terms at this preliminary stage.) Testing the causal thesis is a matter of ascertaining whether certain factors causally influence a subject's performance on this "objectivity scale"—in particular, whether emotional arousal has a causal impact.

It seems reasonable to assume that both elements of the causal connection that we wish to investigate—the subject's level of affective arousal and his/her score on the "objectivity scale"—are continuous phenomena. One can, for example, be emotionally aroused not at all, a little, a fair amount, a great deal. We might expect something similar regarding the strength of objectivity with which a person's moral experiences are imbued. This being so, evidence for a causal connection can be gained via manipulating the hypothesized causal antecedent and observing proportional change in the hypothesized causal consequent. Arousing (certain?) emotions should "ramp up" the subject's tendency to imbue her moral judgments with objectivity; dampening emotions should be accompanied by a reduced experience of objectivity. (Naturally, standard procedures of randomization, and so on, should be enforced.) Of course, this sounds all very easy in principle; no doubt designing adequate experimental protocols will be a far more complicated exercise. One reason I claimed that this might be "more straightforward" than testing the phenomenological thesis (assuming, of course, that the latter thesis has already been tested) is that we know various ways of manipulating subjects' affective attitudes. Studies in the psychology literature that involve arousing certain emotions in subjects (both openly or surreptitiously) are too numerous to require citing.

Regarding both the phenomenological thesis and the causal, it would be naïve to think that there is any one test that might provide confirmation. In both cases, what we should be seeking is experimental "triangulation," whereby we come at the target hypothesis from numerous experimental directions.

Moral Projectivism: The General and the Particular

My principal claim is that confirmation of both the aforementioned sub-theses would amount to an empirical confirmation of minimal projectivism. But would it be a confirmation of moral projectivism *simpliciter*? That depends on two things. The first we have already discussed: There are conceptions of moral projectivism that require the

satisfaction of further sub-theses. If our interest lies in one of these other non-minimal versions of projectivism then we should, of course, still be highly interested in the empirical prospects of minimal projectivism, for its confirmation would count as the confirmation of *a necessary part* of our preferred theory. The second complicating factor—the one I earlier postponed the discussion of—is that it is not clear *how many* token episodes of moral judgment the two sub-theses need hold true of before we can legitimately speak of minimal moral projectivism *as a general thesis* holding true. Let me explain.

It seems to me fair to assume that, once basic determinacy and coherence have been accorded to the thesis of minimal moral projectivism, most people will agree that it holds *sometimes*. To the extent that a psychopathological notion of "projecting one's emotions" is present in vernacular conversation ("He thought that everyone was criticizing him, but really he was just projecting his own insecurities"), it seems plausible to assume that we are generally comfortable with the idea that *sometimes* moral judgments are the result of individuals projecting their emotions onto their experience of social interaction. Yet one can accept this without thereby counting oneself an advocate of *moral projectivism*. So it seems that a reasonable question to ask is "How frequently would the minimal projectivist account of token moral judgments have to hold before we would claim that Minimal Moral Projectivism is in general true?" (I will now use upper case to indicate the general thesis.)

It seems doubtful that the answer should be "Always." Think, by comparison, of projectivism about color (bearing in mind Hume's apparent like treatment of color and morality). The color projectivist need not claim that *every* color judgment is the product of an episode of perceptual projection. If I inform you that my screensaver is predominantly the same color as the sky on a clear day, then you can—without ever laying eyes on my computer—make the judgment that my screensaver is predominantly light blue. One might balk at calling this a "color judgment," but I have no qualms in that respect. You have the concept *blue*. You come to believe that a particular item (my screensaver pattern) falls within the extension of the predicate "… is blue." You may then assert the sentence "The screensaver is blue" and thereby say something true.[21] That sounds like a color judgment to me.

There are many differences between making color judgments on the basis of visual apprehension and on the basis of inference. I guess one obvious difference might be sheer *frequency:* Inferential color judgments seem fairly unusual—at least if we are talking about ones that are explicitly represented in deliberation. This observation might lead one to say that Color Projectivism will be true so long as the projectivist story holds good of *most* color judgments. The fact that projectivism does not hold true of inferential color judgments might be dismissed as statistically insignificant.

But this in fact does not seem to me like the correct way of thinking about the difference. The more salient difference between the two kinds of color judgment is that there

[21] At least: something that has as good a claim to being true as if you were to assert the same sentence on the basis of visual acquaintance with my screensaver.

is a kind of logical asymmetry between them. The inferential color judgment is parasitic on the perceptual color judgment, in the sense that if someone has *never* made a perceptual color judgment (that is, has never experienced color) doubt arises as to whether she can even make an inferential color judgment. Consider a person totally colorblind from birth. Such a person can learn that the sky lies within the extension of the predicate "… is blue," and therefore when it is put to him that my screensaver lies within the extension of the same color predicate, it is a simple matter for him to come to the conclusion that asserting the sentence "The screensaver is blue" will meet with agreement. But does such a person really understand what he is saying? Does he have the concept *blue* at all? Does he have any beliefs about blue things? I do not need to argue that the answer to these questions is ultimately definitively negative; it is enough for my purposes to note that there is at least a temptation to answer them in the negative. (See Peacocke 1983; Tye 1999.)

My point is that there is a way of understanding the asymmetry between inferential color judgments and perceptual color judgments that is not statistical. Even if *most* color judgments were inferential, there would, I suggest, still be this temptation to treat the perceptual color judgments as somehow privileged. One may, then, claim that what it takes for Color Projectivism to be true is for the projectivist story to hold true of all episodes of color judgment *in the privileged class*. (Or, I suppose, one might say that what is required is that it holds true of *most* episodes of color judgment in the privileged class—thereby mixing in something statistical.)

If this sounds along the right lines, then the same strategy should be available to the *moral* projectivist. Can one make a moral judgment without any act of emotional projection involved? "Sure you can," says the Minimal Moral Projectivist. The Minimal Moral Projectivist might claim that this happens frequently, or even usually. What makes him nevertheless a Minimal Moral Projectivist is the conviction that there is a privileged category of moral judgment and the minimal projectivist story is true of all (most of?) the members of that class. Suppose there is some kind of impairment that one might suffer—an imaginary impairment will suffice—that leaves us doubting whether the sufferer really has any moral concepts. The sufferer might be savvy enough to catch on to the socially appropriate sentences to utter—she might know that stealing, promise-breaking, and pedophilia fall within the extension of the predicate "… is morally wrong"—and indeed in our conversations with her we may not even realize that anything is amiss. (We might be similarly fooled by a well-trained but completely colorblind person giving a competent lecture on Impressionism.) But when we discover that the person does in fact have this deficiency—that there is a kind of mental state that she is incapable of having and has never had—we grow doubtful as to whether she really understands what she is saying, whether she has the concept *morally wrong* at all, whether she really has any moral beliefs.[22] If this were so, then the natural thought

[22] The noncognitivist, of course, in a sense denies these things across the board, even for ordinary unimpaired persons. I take it, though, that with a bit of hedging and rewording, the present point about an asymmetry could be expressed in terms amenable even to a noncognitivist.

is that there is an asymmetry in the kinds of moral judgments made by the *unimpaired* persons: that those made in the absence of the mental state(s) in question are parasitic upon those made in its presence. The Minimal Moral Projectivist can then limit his claim to those judgments in the privileged class, irrespective of their statistical frequency.

The question of whether individuals blind from birth have color concepts seems to be an *a priori* one: It is a matter for philosophers to haggle over. It is not so clear that the issue of whether individuals with various kinds of psychological impairment have moral concepts must proceed in an entirely *a priori* matter. Of course, it *might* be that ruminations *au fauteuil* serve to settle the matter; I said above that even an imaginary kind of impairment might be sufficient to ground our conviction that there exists this kind of asymmetry relation. But, on the other hand, it may be that we do not have very strong intuitions on the matter, and that it is only after a course of empirical inquiry that we feel confident in coming to this conclusion. For example, suppose one hears the following:

> Fred suffered brain trauma as a child that left him utterly incapable of feeling empathy. But he often gives normal answers on questionnaires concerning morally loaded vignettes. For example, he assents to the question "Is promise-breaking morally wrong?" Does Fred have the concept *moral wrongness*?

An uninitiated respondent might not have a strong view on the matter. However, it is possible that when we attend to actual cases of impairment, and carefully examine the subjects' responses in a variety of domains (perhaps employing a number of experimental methods), we will acquire a body of data that will lead us to conclude that sufferers of this deficiency lack moral concepts. The obvious cases to look at in this respect are psychopaths and people suffering from various kinds of localized brain damage resulting in what has been dubbed "acquired sociopathy" (see Tranel 1994; Bechara et al. 2000; Ciaramelli et al. 2007).

At the risk of annoying the reader by once more stating what I'm *not* doing in this essay: I don't propose to argue that these kinds of subjects lack moral concepts. My whole point is that this may be a conclusion that we come to only after a careful examination of empirical evidence—much of which may not even be yet gathered. But I will mention that there is already some suggestive data that may point us in this direction. Consider psychopaths. Psychopaths can certainly linguistically respond in an appropriate manner to morally loaded vignettes. (They do not ask "What does this word 'right' mean?" They do not apply the word "right" to utterly inappropriate things, like days of the week or inanimate objects.) And thus they can, at least superficially, demonstrate basic competence with moral terms (as can a blind person competently use color vocabulary). However, psychopaths can also be found to use moral vocabulary in extremely confusing ways in unguarded moments. One such individual, incarcerated for theft, when asked if he had ever committed a violent offense, replied "No, but I once had to kill someone" (Hare 1993: 125). This is not an isolated slip (see Kennett and

Fine 2008 for more examples of this sort of linguistic infelicity in psychopaths). Furthermore, upon more careful examination we discover some strikingly unusual aspects of the psychopath's deeper grasp of morality. Both children and adults with psychopathic tendencies fail to grasp the distinction between moral and conventional norms—a cross-cultural trait that usually emerges at the age of about three. (For references and discussion, see Blair et al. 2005: 57–8.[23]) Experiments also reveal that psychopaths struggle to process certain linguistic information, especially that which is emotionally salient: Whereas normal persons process emotional words faster than neutral words, for psychopaths there is no appreciable difference (Williamson et al. 1991; see also Blair et al. 2005: 59–62 for further references). Psychopaths lack some of the affective input into linguistic processing, and thus, it may well be argued, suffer from a lack of proper understanding of the associated concepts; arguably, they do not qualify as genuinely knowing what a term like "moral goodness" even means. In his classic study of psychopaths in the mid-twentieth century, Hervey Cleckley explicitly likened psychopathy to colorblindness: The psychopath cannot comprehend "goodness, evil, love, horror, and humour . . . It is as though he were colourblind, despite his sharp intelligence, to this aspect of human existence . . . He can repeat the words and say glibly that he understands, and there is no way for him to realize that he does not understand" (1941: 90).[24] After reviewing a number of sources of evidence, Jeanette Kennett and Cordelia Fine conclude that "a growing body of evidence . . . such as their poor performance on the moral-conventional distinction task and their incompetence in the use of evaluative language, suggests that psychopaths deviate so significantly from the folk that it is reasonable, *on empirical grounds*, to conclude that they do not have mastery of the relevant moral concepts" (2008: 219; my italics).

The case may be more complicated than that of the colorblind person's apparent failure to grasp color concepts, since it seems to involve an extra logical step. The first step is to argue by direct analogy with the colorblindness case: If a person has never experienced the emotion of guilt, say, then they cannot really have the concept *guilt*. We might repeat this step for a number of different emotion/affect concepts. The additional step is to argue that grasp of these emotion/affect concepts is a necessary condition for being granted competence with the moral concepts. I am not claiming that either step can be settled exclusively by *a posteriori* inquiry; I am observing that either step may be bolstered and informed by empirical input. A blunt presentation of the question "Fred has no capacity to feel guilt; does he have the concept *guilt*?" may not prompt confident responses. But an examination of the constellation of impairments

[23] Some have expressed doubts about aspects of the moral/conventional distinction (see Kelly et al. 2007; Kelly and Stich 2007), but their skepticism does not extend to casting into doubt the evidence that there exists a substantial performance divergence in this respect in individuals manifesting the psychopathic profile.

[24] Kennett (2002) argues that psychopaths lack the concept *duty* due to their impaired understanding of ends and reasons. Smith (1994: chapter 3) argues that having certain motivations in favor of compliance is necessary for mastery of moral concepts. He draws a direct analogy with how things stand in the case of a colorblind person's grasp of color concepts.

that follow from this emotional incapacity may help settle one's views. Of course, ascertaining which other incapacities truly *follow from* the impairment in question, and which are merely contingently associated with it in the case of psychopathy, is an extremely delicate matter. It is, however, very clearly an extremely delicate *empirical* matter.

In sum: Empirical investigations can contribute much to our deliberations concerning the psychological prerequisites for moral competency. Any such conclusion can then allow us to identify a privileged class of moral judgments, which opens the possibility of embracing a non-statistical notion of what it takes for Minimal Moral Projectivism to be true: It is true so long as the minimal projectivist account holds of members of the privileged class; it is not a matter of *how many* token episodes of moral judgment the minimal projectivist account accurately describes, but *which* episodes.

Conclusion

In this essay I have attempted to clarify a particular version of moral projectivism and have drawn attention to the fact that it is an empirical thesis. Even if there exist other legitimate versions of projectivism requiring the satisfaction of further criteria, I suggest that at their core will lie the same empirical commitments. The minimal version of moral projectivism that I have discussed is neutral between any of the standard metaethical options—and I would claim this as one of its virtues. One might, then, question whether minimal moral projectivism is *metaethically* interesting at all. Despite the fact that my CV bears the title "metaethicist," I must confess to being unsure about where to draw the lines around the discipline. (I am none too comfortable about even calling it a "discipline.") I am quite prepared to endorse an Institutional Theory of metaethics: It concerns whatever metaethicists decide it concerns. Even if minimal projectivism does not count as a "metaethical thesis," the fact that a proposal that has traditionally been thought of as a metaethical thesis turns out not to be one is itself something of which metaethicists should take note. Similarly, it might be claimed that by homing in on a specifically empirical and metaethically neutral version of projectivism, I am stripping the thesis of its *philosophical* interest. I remain blasé, finding any dispute over whether something counts as philosophically interesting extraordinarily philosophically uninteresting.

11

Moral Fictionalism

> Were I not afraid of appearing too philosophical, I should remind my reader of that famous doctrine, supposed to be fully proved in modern times, "That tastes and colours, and all other sensible qualities, lie not in the bodies, but merely in the senses." The case is the same with beauty and deformity, virtue and vice. This doctrine, however, takes off no more from the reality of the latter qualities, than from that of the former; nor need it give any umbrage either to critics or moralists. Though colours were allowed to lie only in the eye, would dyers or painters ever be less regarded or esteemed? There is a sufficient uniformity in the senses and feelings of mankind, to make all these qualities the objects of art and reasoning, and to have the greatest influence on life and manners. And as it is certain, that the discovery above-mentioned in natural philosophy, makes no alteration on action and conduct; why should a like discovery in moral philosophy make any alteration?
>
> David Hume, "The Sceptic" (1742)

If there is Nothing that We Morally Ought to Do, then What Ought We to Do?

On the very last page of his book *Ethics: Inventing Right and Wrong*, John Mackie (1977) suggests that moral discourse—which he has argued is deeply error-laden—can continue with the status of a "useful fiction." I presume that most people will agree, for a variety of reasons, that morality is in some manner useful. The problem, though, is that its usefulness may depend upon its being *believed*, but if we have read the earlier stages of Mackie's book and have been convinced by his arguments, then surely the possibility of believing in morality is no longer an option. Even if we somehow *could* carry on believing in it, surely we should not, for any recommendation in favor of having false beliefs while, at some level, knowing that they are false, is unlikely to be good advice. So how useful can morality be if we don't believe any of it?

This essay will assume without discussion that Mackie's arguments for a moral error theory are cogent (or, at least, that their conclusion is true). This amounts to assuming two things; first, that moral discourse typically is assertoric (that is, moral judgments

express belief states); second, that moral assertions typically are untrue. Mackie's particular argument holds that the problems of morality revolve around its commitment to Kantian categorical imperatives: Morality requires that there are actions that persons ought to perform regardless of their ends.* But, Mackie argues, such imperatives are indefensible, and therefore morality is flawed. A moral error theorist must hold that the problematic element of morality (categorical imperatives, in Mackie's opinion) is *central* to the discourse, such that any "tidied up" discourse, one with the defective elements extirpated, simply wouldn't count as a *moral* system at all.

There are rich and inventive arguments against Mackie, but here we will suppose them all to fail. The question that this essay addresses is "What, then, ought we to do?" Mackie's answer appears to be "Carry on with morality as a fiction," and it is this possibility that I wish to examine closely. The aim is to understand what such an answer may mean, and to attempt a defense of it. I will call the view to be defended "moral fictionalism." Fictionalism promises to be a way by which we can avoid the situation that Quine so deplored, of employing "philosophical double talk which would repudiate an ontology while simultaneously enjoying its benefits" (Quine 1960: 242). Note that fictionalism is not being suggested as something that is true of our actual moral discourse; rather, it is presented as a stance that we could take toward a subject matter—morality, in this case—if we have become convinced that the subject is hopelessly flawed in some respect, such that we cannot in good conscience carry on as before. In the useful terminology of John Burgess, I am peddling a "revolutionary" not a "hermeneutic" fictionalism (Burgess 1983).[1]

One might think that the question "If a moral error theory is the case, what should we do?" is self-undermining. And so it would be, if it were asking what we *morally* ought to do, but that is not what is being asked. It is just a straightforward, common-or-garden, *practical* "ought." The answer that the question invites will be a hypothetical imperative, and we will assume that whatever arguments have led us to a moral error theory have not threatened hypothetical imperatives. (In other words, to hold a moral error theory is not to hold an error theory for practical normativity in general.) I do not want this issue to depend on any particular view of how we make such practical decisions. Let us just say that when morality is removed from the picture, what is practically called for is a matter of a cost-benefit analysis, where the costs and benefits can be understood liberally as preference satisfactions. By asking what *we* ought to do I am asking how a *group* of persons, who share a variety of broad interests, projects, ends—and who have come to the realization that morality is a bankrupt theory—might best carry on. (Two comments: (1) I wouldn't object if we decided to speak of *informed*

* 2015 addendum: I now see this as a misinterpretation of Mackie's argument. Rather, it is only a certain subset of categorical imperatives—those with aspirations of objectivity (or "institution-transcendence, as I have occasionally called it since)—to which Mackie objects. See the introduction to Joyce and Kirchin (2010) for further explanation.

[1] Burgess's original distinction was between two forms of nominalism: See also Burgess and Rosen (1997). For criticisms of hermeneutic fictionalism, see Stanley (2001).

rather than actual preferences; (2) no assumption is being made that preferences will be selfish in content.)

I will begin by discussing fictionalism in general, outlining how it might be that a person might carry on using a discourse that she has come to see as flawed. It will be useful if initially we avoid the distractions that the particular case of *moral* fictionalism might bring, and so I will begin by discussing an example that in some ways is less controversial: color fictionalism.

Critical Contexts

Suppose that after reading some eighteenth-century philosophers David comes to endorse an error theory about color. We needn't go into the arguments that might lead him to this conclusion, but they probably have something to do with the thought that one of the central platitudes about color is that it is a type of surface property of objects with which humans can have direct acquaintance (for example, with normal eyesight on a sunny day), coupled with the thought that there simply aren't any properties like *that*. In other words, for philosophical reasons he ceases to believe that the world is colored in the way that it appears to be colored, which (further philosophical reasons lead him to think) implies that it is not colored at all. Maybe he is confused in coming to such a conclusion, but that is not the issue.

The issue is: Given that he has come to have this philosophical belief (however confusedly) what happens to all his color discourse? Does he stop saying things like "The grass is green"? If someone asks him what color his mother's eyes are, does he reply that they are no color at all? Does he cease to appreciate sunsets or Impressionist paintings? Does he wear clashing clothes (while denying that anything really clashes with anything)? Of course not. In 99 percent of his life he carries on the same as everyone else. His vision is the same, his utterances about the world are the same, and even what he is thinking while making these utterances is the same. It is only in the philosophy classroom—moreover, only when discussing sensory perception—that when pressed on the question of whether the grass is green David might look uncomfortable, squirm, and say "Well, it's not *really* green—nothing is *really* green." This may seem like an uneasy position for him to be in. Sometimes—99 percent of the time, let's say—he is willing to utter "The grass is green," "The sky is blue," and so on, while at other times—1 percent of the time—he is inclined to deny these very same propositions. Which does he believe?

It seems to me that in this case what he affirms 1 percent of the time determines his beliefs. Why? Because the circumstance in which he denies that the world is colored—the philosophy classroom—is the context within which he is at his most undistracted, reflective, and critical. When one thinks critically, one subjects one's attitudes to careful scrutiny ("Is my acceptance of *p* really justified?"); robust forms of skepticism are given serious consideration; one looks for connections and incoherencies amongst one's attitudes; one forms higher-order attitudes toward one's first-order judgments. It is important to see that this distinction between more critical and less critical contexts is

asymmetric. It's not merely that a person attends to *different* beliefs when doing philosophy than when, say, shopping; nor that she questions everyday thinking when doing philosophy, but equally questions philosophy when shopping. Critical thinking investigates and challenges the presuppositions of ordinary thinking in a way that ordinary thinking does not investigate and challenge the presuppositions of critical thinking. Critical thinking is characterized by a tendency to ask oneself questions like "Am I really justified in accepting that things like shops exist?"—whereas the frame of mind one is in when shopping is *not* characterized by asking "Am I justified in accepting that there is some doubt as to whether shops exist?"

This notion of what a person is disposed to assent to if placed in a critical context must not be read as involving any far-fetched counterfactual idealization. Who can judge what manner of bizarre things one would assent to if given *perfect* powers of reflection and critical thinking? A person's "most critical context" must be fixed in actuality—and the obvious means of achieving this grounding is to stipulate that he must sometimes (at a minimum, at least once) have *actually inhabited* that context, and therein either assented to, or dissented from, the thesis in question. In other words, it would be too bizarre to hold that an individual, who has never given the issue any careful thought whatsoever, but thinks and acts in accordance with theory *T*, does not really believe *T* simply because if he *were* to think carefully about it, he would deny it. But if we add that at some point he *has* adopted a critical perspective and therein sincerely denied *T*, and remains disposed to deny *T* were he again to adopt that perspective, then he disbelieves *T*, regardless of how he may think, act, and speak in less critical perspectives. In David's case, his most critical context is philosophical thought—thus, though he occupies this position only 1 percent of the time, we're supposing, it is his pronouncements therein that reveal his beliefs. The rest of the time he still *has* this skeptical belief, but he is not attending to it. Nevertheless, *all* the time David remains disposed to deny that the world is colored if placed in his most undistracted, reflective, and critical context, thus *all* the time this is what he believes.

Fictive Judgments

This leaves us with the question of how we should describe David's color claims in that 99 percent of his life where he utters propositions (for example, "The grass is green") that he disbelieves. We can begin by reminding ourselves of a more familiar circumstance in which people utter propositions that they disbelieve: story-telling. When I utter the sentence "There once was a goblin who liked jam" as part of telling a story, I am not expressing something that I really believe. If pressed in the appropriately serious way ("You don't *really* believe that there once was a goblin who liked jam, do you?") then I will "step out" of the fiction and deny those very propositions that a moment ago I was apparently affirming.

Some people have argued that sentences concerning fiction ought to be interpreted as containing a tacit story operator, such that they may be treated as true assertions;

thus the sentence "There once was a goblin who liked jam" may be used to express the true proposition "According to Hans Christian Andersen's story, there once was a goblin who liked jam." (See, for example, Lewis 1978.) This is inadequate as a general claim, for it fails to distinguish two different things that we can do with a story: describing the story versus telling the story. When we tell a story we are pretending something: that we are a person who has access to a realm of facts that we are reporting. (We might also partially pretend to be characters in the story, which is why we will speak their parts in a gruff or squeaky voice.) But if every sentence of the story uttered contained an unpronounced fiction operator, then there is no sense to be made of the claim that the storyteller is pretending. (How would one *pretend* that according to Hans Christian Andersen's story there once was a goblin who liked jam?)[2] This is not to deny that on occasions the proposition "According to Hans Christian Andersen's story, there once was a goblin who liked jam" might be expressed elliptically, minus the prefix, but this is not what we are doing when we *tell* the story. On such occasions we are not asserting anything, but *pretending* to assert.

The same distinction can be made regarding skeptical David's color claims. When, in ordinary conversation, he utters the sentence "The grass is green" we could interpret this as a kind of shorthand way of asserting something like "According to the fiction of a colored world, the grass is green" *or* we could interpret him as not asserting anything at all, but rather doing something rather like engaging in a make-believe: pretending to assert that the grass is green. I prefer the latter interpretation. It is true that at the moment of making the utterance it doesn't *seem* to David as if he is participating in an act of pretence, but nor does it seem to him as if he's making an implicit reference to the content of a well-known fiction. The matter will not be settled by asking David what he takes himself to be doing. Unless we force him into the philosophical context where he denies the existence of colors altogether, then asking him in an ordinary context whether he is asserting that the grass is green is likely to meet with an affirmative answer. But *that* claim—"Yes, I am asserting that the grass is green"—may be just another part of the fiction. (A Roald Dahl story, recounting many fantastic events, contains an explicit declaration that the story is not a fiction, but it's *all true*. The declaration of truth is no less part of the make-believe than the rest of the story.)[3] The issue of whether David's everyday utterance "The grass is green" is an assertion about a fiction or a fictional assertion is not an issue about how things feel to him—it is to be settled by philosophers providing an interpretation that construes David's linguistic practices most charitably.

The former interpretation—the "tacit story operator view"—does him no favors. One problem is that it cannot account for the fact that when in a more critical context David will explicitly *overturn* what he earlier claimed—he might say "What I said earlier

[2] Walton (1978) makes a similar point.

[3] Dahl's story is "The Wonderful Story of Henry Sugar," in case you're interested. Balzac's *Le Père Goriot* also famously claims of itself that it is neither a fiction nor a romance, but "all is true."

was, strictly speaking, false." But if what he said earlier concerned the content of the fiction of a colored world, then he does not think it was false at all. A second problem with this interpretation is that it fails to make sense of the ways David might employ a color claim in a logically complex context (see Vision 1994). For example, he might endorse the following argument:

P1 Fresh grass is green.
P2 My lawn is made of fresh grass.
C Therefore, my lawn is green.

But if the first premise is elliptical for "According to the fiction of a colored world, fresh grass is green," then the argument is not valid at all. There is room for maintaining that the argument would be valid if all three claims were so prefixed, but the problem then would be that the revised second premise ("According to the fiction of a colored world, my lawn is made of fresh grass") seems so obviously false that it is surely not what David asserts when he utters P2. The fiction of a colored world, insofar as it has a determinate content at all, does not include claims about what anybody's lawn is made of (see comments by Lewis 1978: 38–9).

To this it might be objected that the operator is being interpreted incorrectly. If "according to . . . " means not "it is claimed by . . . " but something more like "it is true in the fiction of . . . " then perhaps we might after all allow that according to the fiction of a colored world my lawn is made of fresh grass. In much the same way we might allow (indeed, insist) that it is true in the fiction of the Conan Doyle stories that humans do not have long hairy tails, that 6 + 5 = 11, that Ireland is to the west of Britain, and so on, despite the fact that one will not find such things *claimed* by the stories (nor even—with, perhaps, the exception of the arithmetical truth—*implied* by anything claimed by the stories).

But this objection leads to unsightly consequences. Suppose David just casually asserts "My lawn is made of fresh grass." Since this assertion may at any time be pressed into service as the premise of an argument (the other premises of which include color claims), if the resulting argument is to be valid we will have to interpret him as *really* having asserted "It is true in the fiction of a colored world that my lawn is made of fresh grass." But the very same assertion may be employed by David as a premise in another argument that involves no color claims and no obvious fictionalizing: He may combine it with "Fresh grass is a type of vegetation," for example, to reach the conclusion "My lawn is made up of a type of vegetation." In order for this new argument to be valid we had better interpret this new premise (and the new conclusion) as also bearing the prefix. In fact, any assertion that David makes might be combined with color claims as a premise of an apparently valid argument, and so if we're to maintain that apparent validity is real validity, we're going to have to interpret everything that he asserts about anything as having this unpronounced prefix. Things become worse still if we remind ourselves that color may not be the only fiction in which David participates. Eighteenth-century philosophy may also lead him to endorse an error theory for

sound and smell, for causation, for virtue and vice, and thus in order for all his apparently unremarkable, apparently valid argumentative moves to be genuinely valid, we will have to interpret every claim issuing from his mouth as brimming with unspoken prefixes.

All such unpleasantness is avoided if we do away with tacit operators, and simply interpret David's utterance "Fresh grass is green" as a kind of make-believe assertion. The content of the proposition doesn't change, any more than when I say (as part of telling a story) "There once was a goblin who liked jam" I am using "jam" with some special meaning. The sentence "There once was a goblin who liked jam" has exactly the same content whether it is used as part of a fairy tale or to foolishly assert something false. What changes is the "force" with which it is uttered. When asserting it I am presenting it as something that I believe, and putting it forward as something that my audience should believe. Linguistic conventions decree that when it has been preceded by "Once upon a time . . . ," all such expectations are lifted.

What are we to make of an argument when some of the premises are uttered as an act of make-believe (for example, as lines in a play) while others are straightforward assertions? Since the presence or absence of assertoric force doesn't affect the content of the premises, then if the argument was valid with its components asserted, it will be valid with them unasserted, and remain valid if some of the components are asserted and some of them are not. For example, the following is a valid argument:

P1 It is cold tonight.
P2 It is the height of summer.
P3 A cold night in the height of summer is unusual weather.
C Tonight is unusual weather.

If a logic teacher recited this argument to a group of incoming undergraduates as an example of validity, she would not be asserting any of the premises or the conclusion—but it would be no less valid for that. Alternatively, suppose that P1 is the line of a play, and the actor duly utters it while on stage, during a performance given on a hot summer's night. After the play, when pressed on climatic issues (curiously), he assents in all seriousness to P2 and P3. Clearly this person has not committed himself to the conclusion (which he may believe to be completely false), for the reason that he did not commit himself to P1. On the other hand, there is nothing to prevent him from "going along" with the pretence if for some reason he wants to, combining P2 and P3 with the make-believe P1, and endorsing the conclusion as part of a fictional act. If he does so, there will be no need to reinterpret his attitude to P2 and P3. These were asserted, and in asserting them he has committed himself to certain other conclusions (for example, "If it were cold tonight, that would be unusual weather"), and may combine them with further asserted premises to yet further conclusions. In other words, unlike with the tacit operator account, we do not have to interpret David's ordinary claim "My lawn is made of fresh grass" as anything other than it appears to be, let alone extravagantly reinterpreting all his other ordinary assertions that are not color claims.

Let us say, then, that David is not only an error theorist about color, but also a fictionalist. He does not believe in color, but he continues to employ color discourse. His color claims are fictive judgments, which we may think of as a kind of "make-believe"—though one should be wary of the term, since the paradigm examples that it tends to bring to mind are of rather trivial activities (pretending that the puppet is talking, make-believing that the sofa is a boat, and so on). But there is no obvious reason to assume that make-believe is always a trivial business;[4] indeed, an important objective of this essay is to convince you otherwise. We have not specified David's reasons for making these fictive color judgments—let us just say that he finds it convenient to do so. This practical value need be nothing more than the convenience of carrying on in the manner to which he has grown accustomed.

Since David is capable of overturning his everyday color discourse whenever he enters a more critical frame of mind, we should hardly describe him as suffering from self-deception. He is no more self-deceived than is someone caught up in a good novel. I suppose that the term "self-deception" *could* be applied to an ordinary person engaged in a novel, but (a) it would be an uncomfortable stretch, and (b) it would merely show that self-deception need not be in the least pernicious.[5] It is much better, I think, to distinguish being "caught up" in a fiction from being "deceived" by a fiction. A person deceived by a fiction is someone who might walk up and down Baker Street wondering where Holmes lived, or who tries to research Madame Bovary's ancestry, or who rushes onto the stage to save the princess. Fans of Sherlock Holmes do travel to Baker Street, of course, and they may well picture their hero there in the nineteenth century, but they know very well (most of them, I hope) what they're doing. At any time, if asked in all seriousness whether Holmes walked these streets, they will answer "No." They are not deceived and therefore not self-deceived; they are merely caught up in a fiction. It is the person who is incapable of dropping the fiction, who continues to speak of Holmes as an historical character even when in her most critical context, who is deceived (though further criteria would need to be met before we would describe such a person as *self*-deceived).

Noncognitivism and the Lone Fictionalist

If by "noncognitivism" we mean the view that a certain discourse does not typically consist of assertions, despite normally coming in the indicative mood, then it would appear that we ought to be noncognitivists about David's fictive color claims.

[4] Autistic children fail to participate properly in games of make-believe, and this corresponds to, and arguably contributes to, a whole range of serious disabilities. See Baron-Cohen (1987); Jarrold et al. (1996). For discussion of the evolutionary importance of make-believe play in humans, see Steen and Owen (2001).

[5] "Self-deception" is a contested term. In this essay I avoid any theoretical commitment on the issue, though I should say that on other occasions I would object to the term being stretched to the extent considered.

Remember that fictionalism is being considered here as something that we could *do* with a problematic discourse, not as an analysis of any actual discourse (problematic or otherwise), thus the same goes for the consequent noncognitive stance: It is a description of a discourse that we might choose to adopt, not a description of an actual discourse. Another thing to note is that although over the years we have grown used to the idea of noncognitivists offering a "translation" of allegedly problematic everyday sentences into some unproblematic idiom, that is not what is being suggested here. For example, we are familiar with moral noncognitivists telling us that a claim like "Stealing is wrong" really amounts to "Stealing: boo!" or "I disapprove of stealing; do so as well!" One might misread the present noncognitivist proposal as suggesting in the same spirit that someone who claims "Stealing is wrong" is really saying something like "Let's pretend that stealing is wrong"—thus making it clear that the claim is not really an assertion. But this would be, as I say, a misreading. When playing a game of make-believe with children—say, crawling around on the floor pretending to be a bear—one might say, in a gruff voice, "I am a bear; I am going to eat you!" It would be an odd theory that identified the true content of this utterance as "Let's pretend that I am a bear; let's pretend that I am going to eat you." Someone saying such things would hardly be "playing a game" at all. He might as well start out saying (in an ordinary voice) "Let's pretend that I am speaking in a gruff voice." With noncognitivism defined as above, it is not incumbent on its proponents to provide a translation scheme from problematic language to unproblematic. For the moral fictionalist/noncognitivist, the content of "Stealing is wrong" is exactly what it appears to be—with whatever erroneous implications she thinks that it has remaining in place. What is different about her utterance of the sentence is the force with which she utters it.

There is, however, a troubling consequence of this kind of noncognitivist proposal, for notice that I claimed that we should be noncognitivists about *David's* fictive color discourse, implying that we might not be noncognitivists about everyone else's color claims. Noncognitivism, thus, becomes a relativistic matter. There is nothing wrong with this per se, but it presents a problem. Does David communicate to other speakers his opinion about the non-existence of color? Unless they discuss matters in a philosophical vein, we can assume not. Thus ordinary speakers will assume that when David utters the sentence "The grass is green" he is expressing a belief. Of course, David could avoid this by employing some of the standard devices for indicating the withdrawal of assertoric force. He could precede his color claims by something equivalent to "Once upon a time . . . "; he could utter them in a sarcastic tone of voice, or in the subjunctive mood; at a pinch, he could wear a T-shirt that declares "I withhold assertoric force from color claims!" But if he does none of these things we can assume that his interlocutors will reasonably take his color utterances to be color assertions. And the possibility arises that if all listeners take an utterance to be an assertion, then, regardless of the speaker's true attitude, it *is* an assertion—in which case maybe we ought not be noncognitivists about David's color discourse after all.

If to assert *p* is to express one's belief that *p*, then it may seem impossible that David could assert "The grass is green," given our assumption that he does not believe this. But this would reveal a misunderstanding of how "express" is intended here: It indicates not a causal relation, but one established by linguistic convention. When one *lies*, for example, one expresses a belief that one does not have. That is to say, one exploits the linguistic conventions that decree that when "Such-and-such" is uttered in certain circumstances (for example, in a serious tone of voice, not as part of a play, not preceded by "Once upon a time . . . " and so on), then the speaker is to be taken to believe that such-and-such. Since, we are assuming, David is not employing any of the well-entrenched devices to indicate withdrawal of assertoric force, then it might be argued that his utterance satisfies the criteria for being an assertion. And since David doesn't believe the proposition in question, then, according to this line of thinking, his alleged assertion that the grass is green looks suspiciously like a *lie*.

It would be nice to avoid the conclusion that fictionalists are liars. Let me offer two responses. First, the term "lie" is a bit steep for the situation described. David, after all, does not intend to deceive anyone when he utters "The grass is green." He has no malevolent agenda. He remains disposed to admit his non-belief in colors if anyone wishes to pursue the philosophical point—it is just that such a cerebral turn is inappropriate for 99 percent of conversations. Although David and his interlocutors may not be on quite the same wavelength when they discuss the color of things, no harm comes of it. If "the truth about David" were to become widely known, then ordinary people may be puzzled or amused at so esoteric an idea as that the world is not colored, but it seems unlikely that they would feel annoyed at having been duped. These comments can be interpreted in either of two ways—I don't mind which: (A) expressing the belief that *p* while not believing that *p* is a necessary but not sufficient condition for lying; or (B) expressing the belief that *p* while not believing that *p* may be a sufficient condition for lying, but lying need not warrant criticism.

The second response is to move attention away from the "lone fictionalist," and remind ourselves that fictionalism is a proposed response to the question of what *we* could do if faced with an error theory concerning a hitherto fully endorsed discourse. Fictionalism may be a stable and viable strategy for a group, even if there are some unsettling aspects of it as an individual stance. A group may have a convention in place that when a certain subject matter is entered into, there is a withdrawal of ordinary conversational force. The question of how such conventions become established and passed on is an intriguing one. Consider the murky origins of the convention of sarcasm, for example. Who decided that a certain tone of voice would act as a kind of derogatory negation of manifest content? We employ the convention without even thinking of it as "a convention"; we do not need to be explicitly taught sarcasm as children, we would have trouble articulating exactly how it works if asked to explain. The convention can also withstand the existence of a sizable number of people in the population who seem oblivious of its existence.

When fictionalism is presented in this light—as a proposal for how a *group* might respond to an error theory—we see just how "revolutionary" are the theory's aspirations. Whether such a radically prescriptive spirit is seen as simply preposterous depends on how we conceive of our philosophical objectives. Do I really expect that ordinary speakers will adjust their attitude toward a problematic discourse? Of course not. Ordinary speakers will carry on doing whatever they please. Most of them believe in ghosts, miracles, astrology, and alien abductions. As philosophers writing against such silly beliefs we conceive of ourselves as correcting erroneous thought—as encouraging people to drop their false beliefs and adopt true ones—but we should not seriously expect to succeed! Revolutionary fictionalism is hardly more ambitious in its prescriptive spirit than this.

The Value of Morality

With a basic theory of fictionalism now on the table, we can turn, finally, to *moral* fictionalism. Suppose that a moral error theory is the case—or at least suppose that a group of people has become convinced of this—what should they do with their faulty moral talk? The conclusion that they should just abolish it, that it should go the way of witch discourse and phlogiston discourse, is certainly a tempting possibility, and may, for all I say here, turn out to be the correct response. But fictionalism shows us that it is not the *only* response; it is at least possible that they may reasonably elect to maintain moral discourse as a fiction. What they need to perform is a cost-benefit analysis. Let us suppose, firstly, that the option of carrying on *believing* in morality is closed to them. They have seen the cat out of the bag and they cannot believe otherwise. Even if they *could* somehow bring themselves sincerely to "forget" that they ever read Mackie's book (for example), surely to embark on such a course is likely to bring negative consequences. I will assume without presenting any arguments that these consequences are sufficiently detrimental as to place this option beyond contention.

Similarly, I will not give serious consideration to the proposal we might call "propagandism": that *some* people may be "in the know" about the moral error theory while, for the greater good, keeping it quiet and encouraging the *hoi polloi* to continue with their sincere (false) moral beliefs. Such a situation really would amount to the promulgation of manipulative lies, which, I will assume, leads ultimately to no good. Here I agree with Richard Garner, commenting on Plato's state policy of deception in the *Republic*:

> If the members of any society should come to believe Socrates' fable [the "myth of the metals"], *or any similarly fabricated radical fiction*, the result would be a very confused group of people, unsure of what to believe, and unable to trust their normal belief-producing mechanisms. It is not wise to risk having a society of epistemological wrecks in order to achieve some projected good through massive deception. (Garner 1993: 96)

Two options remain as contenders in the cost-benefit analysis: abolitionism (or we may call it "eliminativism") and fictionalism. For moral fictionalism to be viable it must

win this pragmatic comparison. It is not required that taking a fictional stance toward moral discourse will supply *all* the benefits that came with sincere moral belief. It can be conceded up front that the pragmatically optimal situation for a group of people to be in is to have the attitude of sincere belief toward moral matters. But it must also be grasped that having a doxastic policy concordant with *critical inquiry* is almost guaranteed to serve better in practical terms for a group than any other policy. We are imagining a group of people whose careful pursuit of truth has overthrown their moral beliefs. Perhaps such people correctly recognize that they were happier and better off before the pursuit brought them so far, but there is now no going back, and to sacrifice the value of critical inquiry would be disastrous.

In order to assess who might win this two-horse race, we must ask the question "What is the value of morality?" Unless we roughly know the answer we can have no idea of what costs its abolition may incur. Let us at first put fictionalism aside, and address the question of the value of morality *when it is believed*. We may then assume that this is a benefit that, *ceteris paribus*, will be lost if a group were to abolish morality, which puts us in a position to ask (in the next section of this essay) whether their adopting a fictionalist stance would allow them to avoid some of those losses.

The popular thought that without morality all hell would break loose in human society is a naive one. Across a vast range of situations we all have perfectly good *prudential* reasons for continuing to act in cooperative ways with our fellow humans. In many situations reciprocal and cooperative relationships bring ongoing rewards to all parties, and do so *a fortiori* when defective behaviors are punished. When, in addition, we factor in the benefits of having a good reputation—a reputation that is based on past performance—then cooperative dispositions can easily out-compete hurtful dispositions on purely egoistic grounds.

To an individual who asks why she should not cheat her fellows if she thinks that she can get away with it, Hobbes long ago provided one kind of answer: because the punishment-enforcing power is very powerful indeed.[6] This answer is developed and supplemented by Hume, who speaks of knaves "betrayed by their own maxims; and while they purpose to cheat with moderation and secrecy, a tempting incident occurs, nature is frail, and they give into the snare; whence they can never extricate themselves, without a total loss of reputation, and the forfeiture of all future trust and confidence with mankind" (Hume [1751] 1998: 82). First, the knave misses out on benefits that by their very nature cannot be gained through defection: "Inward peace of mind,

[6] Given that it is in an individual's interests to engage in mutually beneficial contracts, it will be in her interests to support a social system wherein contractual compliance is enforced. Of course, for any individual the *optimal* scheme is if her neighbors are forced to comply and she alone is able to break contracts and evade punishment—but such an arrangement, we may assume, is not an available option. When the only options concern a *non-discriminating* police force, it will be to each individual's interests to choose the maximally vigilant sovereign power. That way a given individual will have to forego the benefits of cheating others, but stands the best chance of avoiding the proportionally greater costs of being cheated (bearing in mind that the disadvantages of having one's throat cut are far greater than any advantages that may accrue from cutting another's throat).

consciousness of integrity, a satisfactory review of [her] own conduct" ([1751] 1998: 82)—advantages that are constituted by a disposition not to cheat one's fellows. Moreover, the knave will lose these benefits for comparatively trivial gains ("the feverish, empty amusements of luxury and expence"). Third, knaves will be epistemically fallible, and might think that they can get away with something when in fact they will be caught and punished. Fourth, since knaves have on their minds the possibility of cheating whenever they are confident of evading detection, they are likely to be tempted to cheat in situations where the chances of evading detection are less than certain, thus, again, risking severe punishment.

One result we can draw from Hobbes and Hume is that a person may have many reasons for acting in accordance with a moral requirement: the fear of punishment, the desire for an ongoing beneficial relationship, the motivation to maintain a good reputation, the simple fact that one on the whole *likes* one's fellows, that one has been brought up such that acting otherwise makes one feel rotten—all these being solid prudential reasons—plus the moral requirement to act. To subtract the last one leaves the others still very much in play. But if this is so, then what useful role does the last kind of consideration play at all? To answer this it is worth underlining the reference to *temptation* in Hume's answer to the sensible knave. Merely to believe of some action "This is the one that is in my long-term best interests" simply doesn't do the job. Most of us know this from personal experience, but there is abundant empirical evidence available for the dubious (see Ainslie 1975; Schelling 1980; Elster 1984, 1985). Because short-term profit is tangible and present whereas long-term profit is distant and faint, the lure of the immediate may subvert the agent's ability to deliberate properly so as to obtain a valuable delayed benefit, leading him to "rationalize" a poor choice. Hobbes lamented this "perverse desire for present profit" (Hobbes [1642] 1983: 72)—something which Hume blamed for "all dissoluteness and disorder, repentance and misery" (Hume [1751] 1998: 51), adding that a person should embrace "any expedient, by which he may impose a restraint upon himself, and guard against this weakness" (Hume [1740] 1978: 536–7).[7] Let me hypothesize that an important value of moral beliefs is that they function as just such an expedient: supplementing and reinforcing the outputs of prudential reasoning. When a person believes that the valued action is *morally* required—that it *must* be performed whether he likes it or not—then the possibilities for rationalization diminish. If a person believes the action to be required by an authority from which he cannot escape, if he imbues it with a "must-be-doneness" (the categorical element of morality that Mackie found so troublesome), if he believes that in not performing he will not merely frustrate himself, but will become reprehensible and deserving of disapprobation—then he is more likely to perform the action. The distinctive value of categorical imperatives is that they silence calculation, which is a valuable thing when interfering forces can so easily hijack our prudential

[7] I have altered Hume's text from the first person to the third person singular.

calculations. In this manner, moral beliefs function to bolster self-control against practical irrationality.

I would not go so far as to claim that this is *the* value of moral belief, or even the most important benefit—but the argument requires only that we locate one general and reliable source of practical value. This suffices to show why a moral error theorist should hesitate before embracing abolitionism, for it reveals a practical cost that would be incurred on that path. (If there are other sources of practical benefit brought by moral beliefs, then the costs of abolitionism are even higher.) The crucial question, then, is whether some of the costs may be avoided by taking a fictionalist stance toward morality—whether the practical benefits of moral belief may still be gained by an attitude that falls short of belief. On the face of it, it seems unlikely. How can a fiction have the kind of practical impact—moreover, the kind of practical *authority*—that confers on moral belief its instrumental value? This is the major reason that moral fictionalism seems troubling in a way that color fictionalism does not: It seems implausible that a mere fiction could or should have such practical influence on important real-life decisions. In what remains of this essay let me try to assuage this reasonable doubt.

Moral Fictionalism

First let me reiterate the caution already noted: that it is not incumbent on the moral fictionalist to argue that taking a fictional attitude toward morality makes *no* difference, or that morality as a fiction will supply *all* the practical benefits of a believed morality. A background assumption is that the arguments for moral error theory have put the option of a *believed* morality out of the running, so the only comparison in which we are interested is between fictionalism and abolitionism. The fictionalist wins the argument if she shows that there is *some* benefit to be had from keeping moral discourse as a fiction that would be lost (with no compensating gain) by eliminating moral discourse entirely.

In the previous section I argued that an important practical benefit to the individual of having moral beliefs is that they will serve as a bulwark against weakness of will—silencing certain kinds of vulnerable calculation, and thus blocking the temporary re-evaluation of outcomes that is characteristic of short-sighted rationalization. So our task is limited to addressing the question of whether a "mere fiction" could also provide a similar benefit.

A quick argument to show that a positive answer is within reach begins by noting that engagement with fiction can affect our emotional states. This view is not without detractors: Kendall Walton, for example, has argued that fictions do not produce real emotions, but rather make-believe emotions (see his 1978, 1990).[8] But this is a terribly counter-intuitive view, which I am confident is incorrect. All the empirical evidence

[8] Others who reject the view that we have genuine emotions in response to fiction include Kenny (1964) and Budd (1985).

supports common sense on this matter: watching movies, reading novels, or simply engaging one's imagination can produce real episodes of fear, sadness, disgust, anger, and so on. (One explanation is, in the words of two eminent psychologists, simply "that the cognitive evaluations that engender emotions are sufficiently crude that they contain no reality check" (Johnson-Laird and Oatley 2000: 465); alternatively, one may think that the human tendency to enjoy fictional engagement served some adaptive purpose in the ancestral environment.)[9] To this premise we can add the truism that emotional states can affect motivations, and thus behavior. Of course, the emotions arising from fictions do not necessarily affect behavior in the same manner as emotions arising in response to beliefs: The fear of fictional vampires is consistent with my sitting eating popcorn, whereas fear of vampires in which I *believed* would result in purchasing wooden stakes and a lot of garlic. But it does not follow that the emotions arising from engagement with fiction are "motivationally inert." Reading *Anna Karenina* may encourage a person to abandon a doomed love affair; watching *The Blair Witch Project* may lead one to cancel the planned camping trip in the woods. Needless to say, these are not the kind of beneficial behavioral responses that the moral fictionalist is seeking, but they at least show that the causal links between involvement with a fiction and action are undeniably in place.

Let us turn our sights more directly on the question of how a person combats weakness of will. Suppose I am determined to exercise regularly, after a lifetime of lethargy, but find myself succumbing to temptation. An effective strategy will be for me to lay down a strong and authoritative rule: *I must do fifty sit-ups every day, no less*. I am attempting to form a habit, and habits are formed—and, for the doggedly weak of will, maintained—by strictness and overcompensation. Perhaps in truth it doesn't much matter that I do fifty sit-ups every day, so long as I do more-or-less fifty on most days. But by allowing myself the occasional lapse, by giving myself permission *sometimes* to stray from the routine, I pave the way for akratic sabotage of my calculations—I threaten even my doing more-or-less fifty sit-ups on most days. I do better if I encourage myself to think in terms of fifty daily sit-ups as a non-negotiable value, as something I *must* do if I am ever to get fit.

However, to believe sincerely that fifty daily sit-ups are needed in order for me to achieve fitness is to have a false belief (we'll assume), the holding of which will require other compensating false beliefs. If it is true that *more-or-less* fifty sit-ups *nearly* every day is sufficient for health, then that is what I ought to believe. On the other hand, to *pay attention* to this belief exposes me to self-subversion—a slippery slope to inactivity. This is precisely a case where my best interests are served by rehearsing thoughts

[9] The latter hypothesis gains support over the former when one considers that in fictional encounters people enjoy and seek out emotions that they otherwise generally avoid (fear, sadness, and so on). The evolutionary hypothesis holds that the capacity to engage with fiction and make-believe is a kind of "safe training" for real-life risks and opportunities. Natural selection makes the accompanying emotions enjoyable in order to motivate the activity (for the same reason as it makes eating and sex enjoyable). See Steen and Owen (2001).

that are false, and that I know are false, in order to fend off my own weaknesses. But in order to get the benefit from this strategy there is no necessity that I *believe* the thoughts, or attempt to justify them as true when placed in a philosophically critical context. While doing my sit-ups I think to myself "Must . . . do . . . fifty!" but if, on some other occasion, you ask me whether I really *must* do fifty, then I will say "No, sometimes forty would suffice."

Human motivation is often aroused more effectively by mental images than by careful calculation. Hume uses the example of a drunkard "who has seen his companion die of a debauch, and dreads a like accident for himself: but as the memory of it decays away by degrees, his former security returns, and the danger seems less certain and real" (Hume [1740] 1978: 144). Hume's point is that humans put weight on near, recent, and concrete evidence, though there is no rational justification for our doing so. We can imagine the drunkard being presented with impressive statistics on the probabilities of alcoholics suffering an unpleasant end, but remaining quite unmoved; yet one friend dies and he becomes a teetotaler (at least for a while). It's not that he disbelieved the statistics, and the death of the friend need not alter his beliefs about how likely he is to suffer a similar fate, but the "tangibility" of the one death has, in Hume's words, "a superior influence on the judgment, as well as on the passions" (Hume [1740] 1978: 143–4).

If the drunkard has decided that his long-term interests are best served by abstinence, what strategy should he pursue to that end? He should read the statistics, yes, but—perhaps even more importantly—he should attempt to keep the image of his dying friend vivid. He does still better if he can relate that image to his own plight, if he thinks: "If I drink, that's what will happen to me." Now, that proposition is false. What is true is something like "If I drink, there's a 10 percent chance [say] of that happening to me." But *that* thought looks dangerous. He does better with the stronger: "If I drink, that's what *will* happen to me." Yet does he, need he, *believe* this? No: He need not believe it in order for it to affect his actions in the desirable way, and, moreover, he *ought not* to believe it because it is false.

Hume's view that decisions are influenced by the "tangibility" of how information is presented receives ample empirical support. In a large-scale survey conducted on doctors' attitudes toward smoking in the 1970s, it was noted that smoking had dropped most dramatically in chest physicians and radiologists—those who had been exposed to the effects of the activity—while other types of doctor, though no doubt aware of the statistics, were much less moved (Borgida and Nisbett 1977). "Tangibility" also affects the willingness of a person to enter into a mutually beneficial cooperative relationship. It has been shown that pairs of people playing iterated Prisoner Dilemma games will be much more likely to develop a cooperative strategy if the information concerning how the other player acted in the previous round is conveyed by a written note passed through a slot, as opposed to one of two small lights being activated (Enzle et al. 1975). The same information is disclosed by either means, but one form is (in a way that is difficult to articulate) more "concrete," more "palpable," than the other, according it a greater influence in deliberations.

In another study of how people play Prisoner's Dilemma games it was shown that if, while sitting in the waiting room prior to playing the game, a person overhears a (fake) radio news item about an act of sacrifice (such as the donation of a kidney) then the person will be much more likely to adopt a cooperative strategy in the subsequent game (Hornstein et al. 1975). By comparison, a radio story presenting violence and nastiness will encourage listeners subsequently to adopt a non-cooperative strategy. It is possible that a "nice" news story affects the person's mood in a way conducive to cooperation, or perhaps it places in his short-term memory a kind of role model, or temporarily makes certain features of the real world appear more salient in deliberations. However it works, it is pretty clear that an engagement with a fictional story (as opposed to an apparent news item) may have a similar affect (though, to my knowledge, the obvious experiment has not been done).

Although these studies may be unfamiliar, what they reveal should hardly come as a surprise. The whole advertising industry (with which we are all far more familiar than we would wish) operates on the assumption that heavily exaggerated, idealized, and fictional images and narratives can influence real choice. We are shown an image of an absurdly happy family living in an eternally sunny world, and the basis of their rapture, we are encouraged to think, is the cereal that sits in the center of the breakfast table. Do we believe such garbage? Not for a second.[10] Do we, nevertheless, go out and spend our hard-earned money on that cereal? Much as we would like to deny it, masses of empirical research shows that we do.

One may object that choosing breakfast cereals hardly compares to moral decision-making, but it would be naive to deny that the same advertising strategies can encourage us to give to charity, vote for a president, support a bombing campaign, or sign up to join the armed forces. That engagement with fiction can influence our deliberations over the most weighty decisions is beyond question. What is perhaps unusual about the situation of the fictionalist, and which requires more discussion, is the proposal that the action-guiding fiction be in some manner self-generated.

Moral Fictionalism as a Precommitment

Sometimes, when on a long airplane flight, I succumb to weakness of will and eat all the awful in-flight food that I had promised myself I wouldn't eat. It happens because I am trapped and bored with the food right in front of me for a long time. In order to avoid this I have developed a strategy for resisting my own imprudence. If I have decided that I really don't want to eat that slice of cheesecake, but suspect that I won't be able to resist picking at it until it is all gone (despite its tasting of plastic), I smear some gravy on top of it. (It raises the eyebrows of the person sitting next to me, but certainly ensures that I won't eat the cheesecake.) In doing this I am, in a very unglamorous

[10] In a study conducted in 1971, it was shown that only 12 percent of sixth-graders believed that television commercials told the truth all or most of the time (Lyle and Hoffman 1971).

way, following the example of Odysseus when he had himself bound to the mast of his ship so as not to give in to the song of the sirens. The circumstance in which he made that decision was one in which he was free of temptation, but he was shrewd enough to anticipate the overthrow of control. Such strategies for combating weakness of will John Elster calls "precommitments" (Elster 1984: 37ff).

The decision to adopt morality as a fiction is best thought of as a kind of precommitment. It is not being suggested that someone enters a shop, is tempted to steal, decides to adopt morality as a fiction, and thus sustains her prudent though faltering decision not to steal. Rather, the resolution to accept the moral point of view is something that occurred in the person's past, and is now an accustomed way of thinking. Its role is that when entering a shop the possibility of stealing doesn't even enter her mind. If a knave were to say to her "Why not steal?" she would answer without hesitation "No!—Stealing is wrong." What goes through her mind may be exactly the same as what goes through the mind of the sincere moral believer—it need not "feel" like make-believe at all (and thus it may have the same influence on behavior as a belief). The difference between the two need only be a *disposition* that the fictionalist has (though is not paying attention to): the disposition to deny that anything is really morally wrong, when placed in her most critical context.[11]

But what if the knave carries on: "But in all seriousness, taking into account philosophical issues, bearing in mind John Mackie's arguments—*why not steal?*" Then, *ex hypothesi*, our fictionalist will "step out" and admit that there is nothing morally wrong with stealing. So does she then stuff her pockets? No! For she still has all those Hobbesian and Humean reasons to refrain from stealing. It is no part of the argument of this essay that moral thinking should be followed if it prescribes actions that we do not have good reasons for performing independently of moral considerations. One would deny this at the price of allowing that morality may serve no purpose to the individual at all. If we embrace the view that a believed morality is useful to the individual, then we must be employing some non-moral standard by which to make this assessment. If (as seems correct) an individual's believing that some available action is morally required increases the probability of his performing that action, then it seems plausible to assume that the usefulness to an individual of moral belief lies at least in part in its increasing the probability of his performing those actions that he judges he morally ought. From these assumptions it follows that such actions were useful to him anyway: that is, that he had a non-moral reason for performing them.

[11] It is worth reminding ourselves that "critical context" is a term of art, and in other vernacular senses of the phrase it is those times when the person is immersed in the fiction that involve more critical thinking. Working out the plot of a complex novel, for example, may involve a great deal of careful thinking, whereas the thought "It's all just a fiction" is a simple matter. Nevertheless, in the sense defined, the latter is the more "critical context" since it questions and challenges the world of the novel. In the same way, though a moral fictionalist will reject moral claims when doing metaethics, this is perfectly consistent with her employment of the moral fiction at other times involving an enormous amount of critical deliberation and careful calculation.

The idea of the precommitment to the moral fiction being a conscious choice that someone makes is an artificial idealization. (In this it differs from pouring gravy on cheesecake.) It is more likely that a person is simply brought up to think in moral terms; the precommitment is put in place by parents. In childhood, such prescriptions may be presented and accepted as items of belief (it is not implausible to hold that the best way to encourage prudent habits is to tell children a few white lies); thus thinking of certain types of action as "morally right" and others as "morally wrong" becomes natural and ingrained. Later, when a broader and more sophisticated understanding is possible, the person may come to see how philosophically troubling is the idea that there really are actions that people *must* perform, irrespective of whether they wish to, regardless of whether it suits their ends—and if convinced by such arguments she becomes a moral error theorist. But these patterns of thought might be now so deeply embedded that in everyday life she carries on employing them—she finds it convenient and effective to do so, and finds that dropping them leaves her feeling vulnerable to temptations which, if pursued, she judges likely to lead to regret. There is, besides, a practical value to be gained simply from the convenience of carrying on in the manner to which she has grown accustomed. She doesn't cease to be a moral error theorist, but she becomes, in addition, a moral fictionalist.

There are no doubt other ways of combating weakness of will. Perhaps some strategies are, taken alone, more effective than adopting a fictive attitude toward the "must-be-doneness" of the optimal option. All that the present argument requires is that adopting a fictionalist stance would provide *some* help in strengthening resolve in addition to any other effective strategies. (Bear in mind also that I am not arguing that acting as a bulwark against temptation is the *only* value of morality, so even if my arguments concerning the contribution that a moral fiction may make in this respect fail to convince, moral fictionalism does not thereby fall flat.) In fact, the preceding argument entails that there is at least one other effective way of combating weakness of will. Why, one might start out wondering, isn't the decision to adopt morality as a fiction subject to weakness of will? If the presence of the shiny money within reach is likely to tempt one to grab it, ignoring the voice of prudence that is warning that this will lead to no good end, then why won't the same lure of short-term profit also incite the immediate abandonment of the moral fiction? The answer I gave is that the moral fiction is a precommitment that can exclude from practical deliberation the entertainment of certain options: All going well, the fictional attitude blocks the temptation to steal from even arising (just as does, all going well, sincere moral belief). But if this answer is reasonable here, then isn't the same kind of answer, the same kind of prudence-reinforcing strategy, available without any fictionalizing entering the picture at all? Why can't a person simply have the precommitment not to steal (plus a precommitment to keep promises, to refrain from initiating violence, and so on)?

It is not clear what it means simply to have "a precommitment not to steal (and so on)." Perhaps it means a *habit* of not stealing, such that a person is brought up so that the thought of stealing simply doesn't enter his mind. Or perhaps it means a habit of

feeling sympathy for fellow humans, such that the prospect of harming them by stealing from them motivates one to refrain from doing so. But though encouraging such habits may be a very good way of fortifying clear-headed instrumental reasoning (which, for Hobbesian and Humean reasons, generally comes down against stealing), my contention is that they would work even more effectively if supplemented with *moralized* thought.

Suppose that a person with no moralized thinking (neither as belief nor fiction) were, despite his voice of prudence properly counseling otherwise, for some reason to steal. Let's assume that he has in place a habit of not stealing, and a habit of feeling sympathy for others' suffering, but nevertheless these habits were not on this occasion strong enough to withstand the temptation of short-term profit. How does he now feel? The fact that he has broken a habit may surprise him. The fact that he has hurt someone that he didn't want to hurt may cause him disappointment and distress. But the important thing is that he can feel no *guilt*, for guilt requires the thought that one has done something *wrong*. With no moral concepts in play, this person does not have access to the thought that he *deserves* to be punished for his action; he regrets, but he cannot repent. His active sympathy may prompt in him a desire to alleviate the victim's suffering (he may even feel a desire to return the stolen goods), but since he has no thought that he *must* do something to make amends, were he to become distracted by other matters, such that his sympathy for the victim fades, then there is nothing to propel his deliberations back to the resolution that "something must be done." In the end, he has just done something out of character that he wishes he hadn't done. "Sympathy," J. Q. Wilson once wrote, "is a fragile and evanescent emotion. It is easily aroused but quickly forgotten; when remembered but not acted upon, its failure to produce action is easily rationalized. The sight of a lost dog or a wounded fledgling can upset us greatly even though we know that the woods are filled with lost and injured animals" (Wilson 1993: 50).

By comparison, the person who can "moralize" her thoughts (either as belief or fiction) will feel differently if on occasion she succumbs to temptation. She can tell herself that she has done something *wrong*, that her action was *unfair*, that she must make amends, that she not only has risked punishment, but also *deserves* it. (In addition, she can judge that other felons deserve punishment too—a thought that was unavailable to our previous non-moral agent.) The fact that these more robust forms of self-recrimination are available to the moral thinker when she does steal strongly suggests that when she is behaving herself her motivation not to steal is more reliable and steadfast than that of her non-moral counterpart. Her deliberations and justifications do not end in the thought "Well, I just don't want to do that," but rather the more vivid and non-negotiable "That would be wrong."

Of course, what ultimately determines whether a person will refrain from stealing is the strength of the desire not to steal compared with the desire to do so. The claim is that the thought "That would be wrong" plays a role in desire-formation and is likely to *strengthen* any desire against stealing that one has as the result of any "non-moralized"

habit. It is true that this thought as a fictive judgment may not play as robust a role in an agent's desiderative life as the thought as a belief, but so long as it reliably pulls *some* weight—so long, that is, as the fictionalist reliably has a pragmatic advantage over the moral eliminativist—then the error theorist is justified in keeping moral discourse as a "useful fiction."

Conclusion

The advice "Maintain moral discourse as a fiction" is not intended to apply necessarily to any agent in any circumstances. It would be unreasonable to expect that it should, especially since the legitimacy of any more authoritative kind of prescription—for example, to the effect that one *must* adopt the moral fiction, irrespective of one's ends or interests—is likely to have been rejected in the prior argument for a moral error theory (the details of which argument this essay has, for obvious reasons, skirted). It is enough if it turns out to be good advice for *us now*: people who are prone to temptation, epistemically fallible, and familiar with moral thinking. I have offered an argument in support of its being good advice, but of course ultimately it is an empirical matter which depends on the ability to assess far-fetched counterfactuals, and I am the first to admit that it may all turn out to be mistaken. It is possible that moral fictionalism deserves a place on the menu of metaethical options while the prescription urged by those of us on the "revolutionary wing" of the theory remains poor advice.

Since this essay has presented no arguments in favor of a moral error theory, discussing the prospects of moral fictionalism may seem premature. I agree that the preferred strategy must always be to do our utmost to show that moral discourse is not really flawed at all—and I dare say that nearly all readers believe this battle still to be worth fighting. But the viability of moral fictionalism should be of more than academic interest even to those who are not error theorists, for I suspect that those eager to repudiate the error-theoretic position often derive their concern in part from worries about what might *happen* if the theory were to become widely accepted as true. It is viewed not merely as counter-intuitive, but as a genuinely threatening and pernicious doctrine. David Brink, for example, once suggested that we should learn to live with whatever "metaphysical queerness" is entailed by moral realism if the only alternative "would undermine the nature of existing normative practices" (Brink 1989: 173). But if this kind of concern is unjustified—as the possibility of moral fictionalism suggests it may be—then the motivation for resisting a moral error theory is in need of re-examination.[12]

[12] This essay is a rewritten and condensed version of chapters 7 and 8 of *The Myth of Morality* (2001, Cambridge: Cambridge University Press). Some passages are taken straight from this book. Thanks to Stuart Brock, Fred Kroon, and Jerry Vision for useful feedback in the course of rewriting.

12

Psychological Fictionalism, and the Threat of Fictionalist Suicide

Introduction

"Eliminativism" is an ambiguous term. When applied to psychological entities, like beliefs and desires, "eliminativism" (as in "eliminative materialism") usually denotes the view that the entities widely referred to do not exist. This position—championed most famously by Paul Churchland (1979, 1981) and Patricia Churchland (1986)—is an error theory regarding psychological entities. In this context, the verb "eliminate" denotes the act of banishing something from our *ontology*; we realize that there is no place for beliefs and desires in our mature world view. It is not the intention of this essay to assess the truth of psychological eliminativism; despite its being a radical and fairly unpopular view, I will simply assume that there are respectable arguments in its favor. (For defense of the view, beyond the Churchlands, see Rorty 1970; Stich 1983; Ramsey et al. 1990; Taylor 1994.)[1]

In other contexts, by contrast, eliminativism is a theoretical option one may or may not choose to adopt after one has embraced the error theory. Here, the verb "eliminate" denotes the act of banishing something from our *language*; we decide that most uses of the terms "belief," "desire," and so on should be dropped.

Thus one might be a psychological eliminativist twice over: One might decide that there are no such things as beliefs and desires, and then one might decide that most uses of these terms should be jettisoned. Let us call these views "ontological eliminativism" and "linguistic eliminativism," respectively. It is so natural to assume that the latter form of eliminativism should accompany the former that many of the classic statements of eliminative materialism fudge the matter.[2] Yet the two can come apart. One

[1] One can be an eliminative materialist about different categories of mental entity. Daniel Dennett (1988), for example, denies the existence not of propositional attitudes but of qualia. Nevertheless, in this essay I focus exclusively on the case against propositional attitudes, and even there mention only beliefs and desires. Speaking of Dennett, it should be noted that the common assumption that his "intentional stance" is a kind of fictionalism is mistaken (Dennett 1987). He is adamant that his psychological instrumentalism allows for the real existence of beliefs and desires, but as "abstracta" rather than neurological events. "Some instrumentalists have endorsed *fictionalism*," he writes, and immediately declares that his instrumentalism is of a different stripe (1987: 72).

[2] The Churchlands' talk of "theory replacement" is indeterminate between (i) replacing one ontological theory with another, and (ii) replacing one linguistic practice with another. The possibility of fictionalism forces one to make this distinction.

can be an eliminativist in the first sense—endorsing an error theory for certain classes of psychological entity—yet resist eliminativism in the second sense—allowing that talk of these entities should be maintained. This combination of views, which may be called "psychological fictionalism," will be explored in this essay. I will delineate a number of different kinds of psychological fictionalism. My goal is not to advocate any of these theories, though the final section of this essay is devoted to rebutting a charge that may be leveled at psychological fictionalism: the threat of fictionalist suicide.

As a preliminary, let me explain what I mean by saying that according to linguistic eliminativism *most* uses of the offending term should be dropped. The Churchlands liken folk psychology to talk of vitalism (Paul Churchland 1981: 71, 89; Patricia Churchland 1981: 100–1)—something about which one should be an error theorist. Yet nobody claims that *all* appearances of the phrase "vitalistic life force" must be dropped from our discourse—even true sentences like "There is no such thing as vitalistic life force" and "Pasteur designed experiments to test whether there is a vitalistic life force." Rather, the linguistic eliminativist about vitalism argues for the abolition of all utterances that commit the speaker to the existence of vitalistic force. Assertion of these sentences do not carry this commitment. Nor does the mere *utterance* of the sentence "Vitalistic life force exists" commit one to the existence of vitalistic force. One might, for example, utter this sentence without assertoric force, if asked to provide an example of a false four-word sentence; or one might say it as a joke, or as a line in a play. As W. V. Quine once put it: "The parent who tells the Cinderella story is no more committed to admitting a fairy godmother and a pumpkin coach into his own ontology than to admitting the story as true" (Quine 1961: 103). The linguistic eliminativist about psychological entities argues for the abolition of all utterances that commit the speaker to the existence of certain psychological entities.

Characterizing linguistic eliminativism in this way does not succeed in distinguishing it from the fictionalist alternative, for the fictionalist about psychological entities also shuns utterances that commit the speaker to the existence of these entities. Where the theories differ is that the fictionalist holds that all those utterances that one would ordinarily think of as committing the speaker to psychological entities in fact do not (or need not) do so, and thus there is no pressure for their abolition. So, for example, both the eliminativist and fictionalist agree that there is nothing amiss with uttering the sentence "Vitalistic life force exists" in the course of reciting a line of a play, for such a context makes no ontological commitment. Where the difference arises is regarding all those conversational contexts where one would usually assume that the utterance of this sentence does evince a commitment to the existence of vitalistic force: that is, seemingly assertoric talk where there is no acting, no joking, no sarcasm, no quoting, and so on. The eliminativist takes such talk at face value and argues for the abolition of any utterance of "Vitalistic life force exists" from such contexts. Such an attitude may be thought of as the orthodox or commonsensical response. (It is, after all, the attitude we do take toward discourse about vitalism.) But the fictionalist takes an unorthodox route, arguing that the commitment here is only apparent. Those contexts that we would ordinarily

think of as assertoric conversation are in fact, the fictionalist thinks, not dissimilar from reciting the lines of a play: They do not ontologically commit the speaker.

Hermeneutic versus Revolutionary Fictionalism

One kind of fictionalist—the hermeneutic fictionalist—presents the theory as a descriptive analysis of actual linguistic practice. Regarding psychological entities like beliefs and desires, such a fictionalist will argue that though an ordinary utterance of, say, "Mary believes that *p*" *appears* to commit the speaker to the existence of beliefs, in fact it doesn't. There is something about the domain of psychological entities, it is argued, that means that utterances involving their reference are best interpreted as governed by commitment-nullifying conventions. Just as we can nullify commitment by adopting an overtly sarcastic tone of voice, so too (it is argued) can we do so by employing certain terms (like "belief" and "desire"). Such a view counts as ontological eliminativism to the extent that it rejects the existence of beliefs and desires,[3] yet it doesn't comfortably count as an error theory. An error theory, we can assume, accuses an epistemic community of widespread *error*, but if speakers were never in the business of ontologically committing themselves to beliefs and desires when they spoke of them, then they could hardly be accused of an error. Where is the mistake in uttering the falsehood "1 + 2 = 5" if one is just kidding?

A view that allows the illustration of some of these distinctions is Gilbert Ryle's. Ryle (1949) argues that utterances like "Mary believes that *p*" function as *inference tickets*—providing license for the speaker to move from one factual statement to another, while not themselves being factual statements. This may be construed as a form of ontological eliminativism, since it denies the existence of beliefs and desires, and it is, moreover, a hermeneutic view, since it purports to describe how psychological language actually functions. Yet it is not an error theory, inasmuch as it rejects that psychological language was ever really in the business of describing the mind in the first place, and thus could hardly be erroneously misdescribing it. Nor should Ryle's view count as a form of fictionalism, since it in no way interprets the attitude that people adopt in employing terms like "belief" and "desire" to be anything like embracing or discussing a fiction. Fictionalism is not just the negative thesis that utterances that appear to make ontological commitments do not do so, but also the positive (though vague) thesis that the role that these utterances do play is substantively similar to the role of familiar fictional discourse.[4]

[3] It is *possible* to be a hermeneutic fictionalist about Xs without being an ontological eliminativist about Xs, but such a view is unusual and hard to motivate. Such a theory holds that Xs really exist, but that whenever we speak of Xs we do not make assertions that commit us to the existence of Xs. An attempt to advocate this view (or one like it) is made by Christopher Jay in his PhD thesis *Realistic Fictionalism* (University College London, submitted 2011).

[4] Another view that might be categorized similarly to Ryle's, given our current taxonomic concerns, is that of Wilfrid Sellars (Paul Churchland's PhD supervisor). Sellars (1956) argues that an utterance of "Mary

Another kind of fictionalist—the revolutionary fictionalist—presents the theory as a recommendation of how we should change our linguistic practices. Such a fictionalist is an ontological eliminativist and an error theorist: accusing ordinary speakers of making false assertions when they say things like "Mary believes that *p*" and "John desires that *q*." This fictionalist argues that, in actual discourse, when people assert these sentences they really are committing themselves to the existence of (nonexistent) beliefs and desires. But whereas the linguistic eliminativist responds that these utterances must therefore be banished, the revolutionary fictionalist responds that we can alter our practices such that the utterances no longer carry the problematic commitment. In fact, the revolutionary fictionalist argues that we *should* alter our linguistic practices in this manner. We should carry on saying false things, but we should stop *asserting* them.

At this point, many questions jostle for attention. How could we take psychological discourse seriously if it lacked commitment to psychological entities? What would be the point of having such a non-committed discourse? Why should altering the ontological commitments of a linguistic practice be superior to dropping the practice? I shan't have space to address all these natural questions here, though I shall have something to say about some of them. Rather, what I want to describe is a further distinction in how the fictionalist might describe the non-committed practice: between cognitivist and noncognitivist versions of fictionalism. This distinction runs obliquely to the aforementioned one, allowing for cognitivist hermeneutic fictionalism, noncognitivist hermeneutic fictionalism, cognitivist revolutionary fictionalism, and noncognitivist revolutionary fictionalism.

Cognitivist versus Noncognitivist Fictionalism

The cognitivist fictionalist holds that sentences like "Mary believes that *p*" should be construed as containing a tacit story operator: "According to fiction *F*, Mary believes that *p*." This view counts as cognitivist in virtue of the fact that it allows that "Mary believes that *p*" may be asserted and may be true; in such circumstances it will be a true assertion *about* a fiction. That ontological commitment to beliefs is nevertheless removed becomes clear when we compare "According to Perrault, a fairy godmother changed a pumpkin into a carriage," the assertion of which carries commitment to the existence of neither godmothers nor transforming pumpkins.

The cognitivist fictionalist owes us an account of what the relevant "fiction *F*" is, for in order to assess the truth of a statement like "Mary believes that *p*," we will need to know what follows and what does not follow "according to fiction *F*."

In the case of a story by Perrault, one might think that we have a pretty good idea as to what the content of the relevant fiction is: We need merely to read the story. Yet it is

believes that *p*" functions not to state an empirical fact but rather to perform a normative act of justification. (Of course, Sellars's view has many subtleties and is open to interpretation.)

not quite so straightforward, for not all things that are true according to the story are stated explicitly therein. It is not explicitly stated that Cinderella has five toes on each foot, yet it is reasonable to assume that it is true in the Perrault story that she does have five toes on each foot. (See Lewis 1978.) However, it is very difficult to assess the extent of the domain of truths that may reasonably be considered true-but-not-explicitly-stated in a fictional story. It is not explicitly stated that Cinderella has forty-six chromosomes, yet is it true in the Perrault story that she has forty-six chromosomes? The fairy tale is set in a world that seems reminiscent of eighteenth-century Europe, yet is it true in the Perrault story that the Roman Empire fell centuries earlier? I, for one, do not have much of an intuition on these questions, and I suspect it is because the matter is indeterminate. Still, generally speaking this indeterminacy in fictions doesn't get us into any trouble, since the indeterminacy resides in exactly the places where it doesn't matter one way or the other to the consumer of the fiction.

In the case of psychological fictionalism, the fiction in question might be called "folk psychology." This is the theory that eliminative materialists think is false. But even if false, the theory presumably has enough content to ground "According to folk psychology…" claims. Some are straightforward: "According to folk psychology, beliefs exist"; "According to folk psychology, some desires are strong and some are weak." Some will be more complex: "According to folk psychology, if S desires X, and S believes that ϕing is the optimal means of attaining X, then this may lead to S's ϕing." (See Churchland 1981: 71.) But what about a claim that concerns a particular, such as "According to folk psychology, Mary believes that *p*"? The theory of folk psychology presumably makes no reference to the individual Mary. The sentence is, nevertheless, a permissible one. Compare the claim "According to Einsteinian physics, the light from star 66 Tauri is deflected by the Sun's gravitational field." Einstein's theory is general; its complete articulation makes no reference to particulars like "66 Tauri" or "the Sun." The claim is nevertheless reasonable in virtue of Einstein's theory *in conjunction with data concerning particulars* predicting or describing the light from a particular star being deflected by the Sun. In a similar way, the theory of folk psychology in conjunction with data concerning particulars will predict or describe certain phenomena, such as Mary's believing that *p*. This data will concern those things that we would ordinarily take to be evidence of Mary's believing that *p*: her behavior, her saying so, the fact that she was just told this and didn't object, and so on. Indeed, it might be a tenet of folk psychology that certain types of things count as evidence of a person's beliefs. When we couple this tenet with the observation that Mary manifests this kind of evidence, it becomes perfectly reasonable to say "According to folk psychology, Mary believes that *p*."

The cognitivist fictionalist, as we have seen, translates the sentence "Mary believes that *p*" into something that is assertible and has truth value. One might instead translate it into something that cannot be asserted and lacks truth value—perhaps something in the cohortative mood, like "Let's pretend that Mary believes that *p*." This would be a kind of *noncognitivist* fictionalism. Another kind of noncognitivist fictionalist offers no translation at all, but rather treats the sentence "Mary believes that *p*" as more

like something uttered in the course of telling a story than an assertion about a story. Story operators are employed when one is explicitly talking *about* a fiction, but when one is *telling* a story one is engaged in a quite different activity. Nothing approximating "According to the Perrault story, a pumpkin turned into a coach" (or "Let's pretend that a pumpkin turned into a coach") is uttered in the course of *telling* Perrault's story. When one tells the story, one does not assert that a pumpkin turned into a coach; one does something more like pretending to assert it (Searle 1975; Lamarque 1981: 332). The analogous psychological noncognitivist fictionalist thinks that we do not (or need not) assert that Mary believes that *p*, but rather we do something more like pretending to assert it. The view counts as noncognitivist in virtue of the fact that it interprets an utterance of "Mary believes that *p*" as having assertoric force withheld. As with the cognitivist fictionalist interpretation, this removes ontological commitment to beliefs (and other problematic mental states).

Unsuspecting Fictionalizing

None of these views in its hermeneutic form is committed to holding that ordinary speakers are consciously aware of what is claimed of their discourse. The cognitivist fictionalist need not maintain that ordinary speakers *think* "According to folk psychology . . . " The noncognitivist fictionalist need not maintain that ordinary speakers *consider* their utterances to be make-believe assertions rather than real assertions. Rather, these are intended as charitable interpretations of their linguistic practices motivated by the ambition of avoiding the error theory. (Compare Harman and Thomson 1996: 4.) Of course, it remains to be seen whether any of these views is adequate. There may be strong reasons against interpreting belief claims as containing a tacit story operator or as make-believe assertions—reasons that outweigh the (supposed) advantages of avoiding error.

The revolutionary fictionalist, by contrast, seems committed to the fictive attitude being something of which the speaker may be aware. After all, the revolutionary fictionalist is recommending a *change* between the prerevolutionary erroneous utterances and the post-revolutionary trouble-free fictionalist utterances. If the fictionalist account were construed as a charitable interpretation that can be made of the post-revolutionary discourse (despite speaker ignorance of the fact), and there is no discernible difference between the prerevolutionary speaker and the post-revolutionary speaker, then there seems nothing to prevent that same charitable interpretation being made of the prerevolutionary discourse—in which case, revolutionary fictionalism would collapse into hermeneutic fictionalism.

However, saying that the difference between belief and make-believe is "something of which the speaker may be aware" is not to say that the speaker is constantly aware of the difference. What typically distinguishes acts of make-believe, for example, is the fact that when push comes to shove they will be abandoned in favor of belief. We can be confident that Mrs Lincoln's enjoyable engagement with the play ceased the moment

John Wilkes Booth so rudely interrupted the evening's entertainment. But the fact that one *will* abandon one's make-believe (in such-and-such circumstances) is a dispositional property that may not have any phenomenological bite. Suppose a speaker goes around uttering sentence *p* in everyday life, but has the disposition to assert *not-p* if asked in an appropriately serious way—a way that takes into consideration skeptical arguments against *p*, and so on. This may suffice for us to conclude that the speaker doesn't wholeheartedly believe that *p*; indeed, we might say that *p* is being treated as a kind of make-believe. That the speaker instantiates this disposition is something of which she *could* be made aware, but it's not something of which we are forced to say she *must* be aware. Therefore one might be engaged in an act of make-believe without being aware of the fact.

A different argument to much the same conclusion focuses on the fact that which speech act a speaker performs is not determined solely by speaker intentions; the conventions of the wider linguistic community also play a major role. Suppose a newcomer to our community is introduced to the word "slut," yet is taught it as if it's a perfectly descriptive term (denoting a woman with multiple sex partners, say) with no pejorative connotations. (This would not be teaching the person competence with the term.) The speaker goes forth and innocently uses the word whenever he needs to refer to a woman with multiple sex partners. When he employs the term on a given occasion, can we say that he succeeds in using it in a purely descriptive fashion, simply because that's his intention? I wouldn't think so. Does he, rather, use the term in a pejorative fashion, despite his intentions to the contrary? Now I am not so sure, but there is at least something to be said in favor of the positive answer. Those to whom he has talked will most naturally report "He said really insulting things," and can reasonably continue to describe matters in this way even when they realize the misunderstanding ("He said really insulting things, even though I now see that he didn't intend to"). Much the same thing can be said, I think, about acts of make-believe. A linguistic community may have a convention in place that decrees that when a particular phrase is used (something like "Once upon a time . . . ") an act of make-believe is entered into. But a token individual may be unaware of this convention (for whatever reason), and utter the phrase without intending to introduce an act of make-believe. Despite this lack of speaker intention, the surrounding conventions may dominate, ruling that what the speaker uttered after this phrase *is* make-believe. If this is correct, then we see again that one might be engaged in an act of make-believe without being aware of the fact; though, again, this is not to deny that any speaker may be brought to be aware of the fact.

The Benefits of Fictionalism

Let us now ask the obvious question: What would be the advantage of taking a fictive attitude toward folk psychology? If the theory of folk psychology, taken at face value, is false, then surely the honest thing to do is simply to declare it as such and eliminate it

from our common practices? The fictionalist answer must be that folk psychology is *useful*, such that the cost of eliminating it is greater than the cost of taking some kind of fictive attitude toward it. (This is not to claim that the benefits of taking a fictive attitude toward folk psychology are as great as the benefits of believing it. But we are assuming here that the arguments for eliminative materialism have refuted the thesis that folk-psychological claims can be treated as true, if taken at face value as evincing ontological commitment to psychological entities.)

We must tread carefully hereabouts, since the claim that folk psychology has *not* pulled its weight in the practical sphere has been often taken as evidence in favor of ontological eliminativism. Paul Churchland says of the history of folk psychology that "the story is one of retreat, infertility, and decadence"—that an assessment of its explanatory success reveals "a very long period of stagnation and infertility" (1981: 74).[5] The psychological fictionalist embraces ontological eliminativism (and thus is presumably persuaded by arguments in its favor) and yet at the same time recommends maintaining folk psychology as a fiction on the grounds of its usefulness. There are several ways of navigating through this apparent tension.

First, the fictionalist may accept that folk psychology has not been sufficiently useful to warrant our thinking it true, but nevertheless allow that it has *some* practical uses—uses that provide its fictional adoption with a practical edge over linguistic eliminativism. After all, the kind of failure to which Churchland draws attention is the failure of a scientific theory (regarding, for example, explanatory success and "coherence and continuity with fertile and well-established theories in adjacent and overlapping domains" (1981: 73)), and he may be correct that such criteria are "the final measure of any hypothesis," such that failure suggests falsehood. But something may disappoint by this criterion yet still be very useful in other ways. Consider the works of Shakespeare.

Second, the fictionalist may have become convinced of the bankruptcy of folk psychology (taken at face value) on grounds other than its practical failure. He or she may accept all along that folk psychology has been and remains a jolly useful theory, but nevertheless have other reasons for thinking it false. This second strategy would be generally untenable only if one endorsed a kind of pragmatism about truth that seeks to exclude the very idea of the *false but useful*. While I have no problem with certain tenets of pragmatism, any version that is so gross as to countenance excluding this idea should have few adherents. (If I point a gun at you and demand that you utter "1 + 2 = 5," does this proposition—suddenly so acutely useful to you—become true (for you)?) Everybody sensible allows the possibility of false propositions whose utterance may be in some manner useful; and to this extent there is an accepted paradigm to which the fictionalist can appeal.[6]

[5] Critics of Churchland on this point include Horgan and Woodward (1985) and Lahav (1992).

[6] See my critical discussion of Hilary Putnam's thinking on this matter (Putnam 1971: 68ff) in Joyce (2001: 189–90).

Relativistic Fictionalism

With talk of usefulness inevitably comes the possibility of relativism, for what is useful to one group or one individual is not useful to another group or individual. I will discuss this matter with respect to hermeneutic and revolutionary fictionalism in turn.

Consider first hermeneutic fictionalism. Why would one prefer a fictionalist construal of the discourse over the more orthodox error theory which follows (we are assuming) if belief and desire claims are taken at face value? The fictionalist's answer is that there is some cost to allowing all utterances of the form "S believes that *p*" (say) to count as false, and so ordinary speakers will instead be interpreted as engaging with a fiction. But suppose that there is some individual—call him "Paul"—for whom this is not true. Paul is unusually situated such that the falsity of all sentences of folk psychology suits him very well. Or we might imagine that Paul is part of a group of whom this is true. Or we might imagine a whole society of whom this is true—not *our* society, presumably, nor even any actual society; but at least a possible society for whom ontological and linguistic eliminativism about folk psychology are no great burden.

It is not clear to what extent we must embrace relativism here: interpreting some speakers' utterances in a fictionalist manner but not other speakers'. What the hermeneutic fictionalist is doing is offering an interpretation of a linguistic *practice*, and if we are to interpret Paul's utterances as part of that same practice, then he may just have to put up with having the same "charitable" interpretation made of his utterances as everyone else, even though it may not suit him personally. In the same way, if some fool happens to think that the word "dog" denotes felines, the fact that it would suit her very well if it *did* denote felines (for then she would triumphantly be proven right) is insufficient to underwrite a relativism about reference such that from her lips "dog" means feline while for everyone else it means canine. Things may stand differently if there were a *group* who started to use "dog" to denote felines. Then they would create a linguistic convention that might demand a special interpretation. (Think of surfers describing an impressive wave as "bad" or "wicked" or "sick.") Perhaps the same would be true of a group whose interests would not be served by rescuing folk-psychological discourse from error; perhaps their utterances of the form "S believes that *p*" (and so on) shouldn't receive the fictionalist interpretation.[7]

[7] The analogy may be importantly misleading. The speaker who misuses "dog" is making a *semantic* error. The question of whether "S believes that *p*" should be interpreted as "According to folk psychology, S believes that *p*" is also, I take it, a matter of the *semantics* of the predicate "… believes that …" However, the question of whether a speaker is, when uttering "S believes that *p*," *asserting it or pretending to assert it* is a matter of the *pragmatics* of the language. Hence the "linguistic conventions" of which I airily speak may be quite different in the different cases, such that the point about an individual's intentions being dominated by the surrounding linguistic population may hold in one case but not the other. I don't have the space to pursue and straighten out the nuances here.

As for imaginary linguistic communities, I suspect that there just isn't a settled truth to the matter of whether we must interpret their utterances relative to what is most useful/charitable *for them*, or give the same interpretation of their utterances as we do of ours, which are interpreted relative to what is most useful/charitable *for us*. Even though there might exist a linguistic community at another possible world that would be better off having its own folk psychology sentences interpreted at face value (thus rendering them false), nevertheless *we* are interested in providing the best interpretation of *our* linguistic practices, and we are entitled to employ that interpretation when considering counterpart discourses at nearby possible worlds. If a fictionalist interpretation (of either a cognitivist or noncognitivist flavor) of actual folk-psychological discourse is warranted on pragmatic grounds, then we will likely impose it on imaginary folk-psychological discourses—even ones for which the cost-benefit analysis works out very differently from our own.

Consider now revolutionary fictionalism. Assuming that an error theory is true of actual folk-psychological discourse, why would one recommend some form of fictionalist response over the more orthodox linguistic eliminativism? The fictionalist's answer is that there is some cost to dropping folk-psychological discourse entirely from our lives, and so ordinary speakers are advised to recoup some of those costs by engaging with a fiction. Here a degree of relativism seems inevitable, for the revolutionary fictionalist is essentially offering *advice*, and what is good advice for one person or group may simply not be for another individual or group, due to their being differently situated. Certainly we can at the very least *imagine* whole populations for whom the fictionalist option is suboptimal. Perhaps these are populations for whom linguistic eliminativism regarding folk psychology is a negligible liability, or perhaps for them the adoption of the fictionalist option brings some large hidden cost. Indeed, it would not be surprising for the fictionalist to think of the fictive attitude as a kind of interim arrangement: as warranted for us now (us for whom folk psychology is an entrenched habit, but who have, let's say, just come to accept ontological eliminativism about psychological entities), but as something that might eventually be discarded as we leave folk psychology entirely behind us. In other words, making a fiction of folk psychology may be good advice relative to us now, but may be poor advice relative to some future Brave New World.[8]

Relativism regarding *individuals* is a more complicated affair, some of the reasons for which have already been touched upon. Suppose Paul is personally ready to embrace linguistic eliminativism about folk psychology: No need for any spineless fictionalist nonsense for him! But let's say that he is unusual; pretty much everyone else would find linguistic eliminativism overly burdensome—more burdensome, at least, than keeping folk psychology alive as a fiction. I see no problem in concluding that revolutionary fictionalism is reasonable relative to most people in this population but

[8] The idea of the fictive attitude (toward a variety of things) as an interim stance that lies between naive belief and enlightened disbelief, is inherent in the work of Hans Vaihinger (1935).

not reasonable relative to Paul.[9] Paul can choose simply not to employ the terms of folk psychology, even though they are being widely used around him. (Think of an atheist in a community of devout theists.) This may be challenging for Paul in practical terms (what's he going to say when asked "Do you believe that it will stop raining soon?"), but perhaps he can find ways of getting by. One complication that this reveals is that when one weighs the costs and benefits of fictionalism versus eliminativism for an individual, what the other individuals are doing in this respect is an important factor. Maybe eliminativism would be optimal for Paul so long as a sufficient number of his fellows join him, but perhaps if he is surrounded entirely by fictionalists then the inconveniences of being constantly misunderstood speak in favor of his taking the same option as them. (In other words, perhaps the question of which is the optimal attitude for an individual to adopt exhibits some of the complexities of frequency-dependent Darwinian selection.)

More complications are revealed if we try to imagine the reverse: an individual for whom making a fiction of folk psychology is optimal, who is surrounded by a sea of linguistic eliminativists. The problem here is that it is not obvious that one can properly make a fiction of a discourse *by oneself*; a linguistic community is needed in order to create the necessary linguistic conventions. Imagine someone in our midst going round talking about phlogiston—pointing at flames and saying "Here is the phlogiston escaping." We would naturally take her to be making badly mistaken assertions, to be embracing a false theory. Could it be, though, that she is just pretending to assert these things, or that she is really making true claims with a tacit story operator ("According to phlogiston theory, here is the phlogiston escaping")? If she makes no effort to inform us what's going on, then it is not clear to me that the answer to either question is "Yes." As I suggested earlier, the conventions of the surrounding linguistic population may dominate the speaker's intentions and determine what speech act is performed. A person may intend to *pretend* to assert X, but if everyone takes her to be really asserting X then arguably she really does assert X. A person who performs a paradigmatically rude act, such as spitting on the carpet, cannot defend himself by saying "I wasn't really being rude; I was just pretending to be." Or spare a thought for the poor genius who invented sarcasm: Imagine him or her going round trying out the new tone of voice on people without explaining how it works.[10]

In sum, the revolutionary fictionalist's recommendation may be reasonable relative to some individuals and unreasonable relative to others, but in some cases the criteria that determine the matter may include that the individual is a member of a like-minded group of sufficient size and structure to underwrite a certain kind of linguistic convention.

[9] Since revolutionary fictionalism is a practical recommendation, the question of relativism must be expressed as "what is *reasonable*, relative to X," rather than "what is *true*, relative to X."

[10] Of course, I don't seriously think that anything remotely like this occurred when the conventions of sarcasm emerged.

Fictionalist Suicide

Many readers will by now be impatient to see a glaring worry with psychological fictionalism addressed: that the thesis is incoherent in that it recommends that we adopt psychological states of which at the same time it doubts the existence. The articulation of the problem starts with ontological eliminativism, which rejects the existence of propositional attitudes like beliefs, desires, hopes, fears, dreads, wishes, and so forth. The cognitivist fictionalist, however, makes essential reference to *belief*: either (a) interpreting psychological claims (such as "Mary desires that *p*") as expressing *beliefs* about what is true according to folk psychology, or (b) recommending that we change our attitude toward such claims such that they come to express such beliefs. The noncognitivist fictionalist, by contrast, makes reference to *make-believe* rather than belief. But it seems plausible to maintain that make-believe is another category of folk psychology (see Currie 1990; Nichols and Stich 2003)—or at the very least that it is a kind of mental state of which we can make sense only within a framework that includes the propositional attitudes of folk psychology—in which case noncognitivist fictionalism also seems to be describing or recommending a kind of mental state whose existence it denies.

This problem of *fictionalist suicide* is reminiscent of a complaint that has been made against ontological psychological eliminativism in general: that it commits *cognitive suicide*. (See Cling 1989; Boghossian 1990; Devitt and Rey 1991; Reppert 1992.) There the criticism is that the ontological eliminativist denies the existence of beliefs, but in articulating and arguing for this thesis presumably puts forward propositions that express his/her *beliefs* and are recommended to the audience as items to be *believed*. In my opinion, the Churchlands responded successfully to this complaint long ago, with an especially effective parody involving the impossibility of anyone coherently denying the existence of vitalistic life force:

> The anti-vitalist says that there is no such thing as vital spirit. This claim is self-refuting; the speaker can expect to be taken seriously only if his claim cannot. For if the claim is true, then the speaker does not have vital spirit, and must be dead. But since dead men tell no tales, they do not tell anti-vitalist ones either. One cannot reason with dead men. (Patricia Churchland 1981: 100)

The argument is obviously question-begging. Someone who denies the existence of vitalistic life force at the same time denies that he (the person advocating the argument) must have vitalistic life force in order to present a meaningful argument. In the same way, the ontological eliminativist about beliefs not only denies their existence but also denies that what goes on when an argument is articulated, advocated, and accepted in any way presupposes the existence of beliefs. What the articulation, advocacy, and acceptance of an argument *does* involve is, quite literally, hard to say. It is hard to say because what it really involves are complex neurological states about which our knowledge is patchy and for which ordinary language is lacking.

The challenge of fictionalist suicide, however, is not so easily overcome as the challenge of cognitive suicide. That ontological eliminativism is put forward as an item to

be *believed* is not itself part of the content nor a presupposition of that theory. By contrast, and more worryingly, that we are able to *make-believe* that a fiction is true *is* an essential presupposition of certain forms of fictionalism. (Likewise, *mutatis mutandis*, with other forms of fictionalism.) This challenge is unique to psychological fictionalism; it doesn't arise for moral fictionalism, or modal fictionalism, or mathematical fictionalism, and so on. This is because the very statement of any fictionalist theory (such as "Make-believe that error-laden discourse X is true") involves reference to psychological entities, but not to moral, modal, or mathematical (and so on) entities.

One way someone might try to overcome this challenge is by denying that the kind of mental states required by fictionalism are among those rejected by the ontological eliminativist.[11] Rather than "make-believe," the key fictive attitude is sometimes called "acceptance"—a kind of embrace that falls short of belief (Van Fraassen 1980). The ontological eliminativist holds that the folk-psychological concept of *belief* indispensably involves a set of criteria that nothing satisfies. (William Alston (1996), for example, lists half a dozen plausible contenders for such criteria.) *Acceptance*, by contrast, might be taken to be something that requires some but not all of these criteria, such that (a) certain actual psychological states do satisfy the criteria for acceptance, and (b) acceptance is not a folk-psychological notion. Thus, one might maintain that acceptance is exactly what should replace the flawed folk-psychological notion of belief. The ontological eliminativist, accordingly, may have no objection to *acceptance*, and is thus free to recommend that we adopt this attitude, or to analyze extant states (erroneously taken to be belief) as being in fact states of acceptance.

The problem with this way of overcoming the challenge of fictionalist suicide is that it would achieve far too much. Someone who interprets those states which would ordinarily be considered *belief* to be instead instances of *acceptance* presumably intends this construal to apply across the board. Thus, such an interpretation wouldn't be limited to the fictionalist's proposal that we accept (but not believe) folk-psychological claims, but seems committed to an infinitude of similar proposals: that we accept (but not believe) that gold has the atomic number 79, that we accept (but not believe) that Napoleon lost the Battle of Waterloo, that we accept (but not believe) that the number 84 succeeds the number 83, and so on. My complaint is not that this is an incoherent (or even particularly implausible) perspective to adopt, but that the fictionalist proposal under consideration—pertaining to the attitude that should be adopted toward folk psychology—would lose identity as a distinct theoretical position.

Let me turn to another way that the psychological fictionalist might try to overcome the challenge.[12] Begin by recalling what I said a little earlier about the puzzle of cognitive suicide. Let's allow that the ontological eliminativist does articulate, advocate, and expect her audience to accept her theory, but she does not thereby undermine her theory since these phenomena (articulation, advocacy, and acceptance) are to be

[11] I thank Stuart Brock for pointing out this possibility to me.
[12] Here I develop thoughts presented by Meg Wallace, in an unpublished manuscript.

understood not as involving beliefs, but as involving neurological events for which we lack natural vocabulary. Thus, the ontological eliminativist may allow that when we would ordinarily take ourselves to have grounds for claiming "Mary believes that *p*" and "John believes that *p*," there is in all probability something similar going on in Mary's brain to what is going on in John's brain. And were we instead to take ourselves to have grounds for asserting "Mary desires that *q*," there will be yet another kind of neurological event occurring in Mary's brain. But (the eliminativist thinks) these neurological states do not deserve the labels "belief" and "desire" (and nor does anything else). (The eliminativist may be mistaken about this—some would prefer to identify these neurological states with the entities of folk psychology—but it is not my intention here to assess the eliminativist's case but rather to provisionally grant it.)[13]

An analogy may help. We are all, I take it, error theorists about witches. Yet it is possible that all and only those women who were accused of witchcraft did have some distinctive set of properties. Perhaps they were disempowered women who all threatened the patriarchal society in a certain manner.[14] Perhaps it was something more disjunctive and complicated than that. At the very least we know that there was one property that all and only women accused of witchcraft instantiated: that of being a woman accused of witchcraft. Thus when people accused certain women of being witches, there was something distinctive going on with these women—there was some property or properties that they and only they instantiated—and yet it nevertheless remains perfectly reasonable to insist that there are no such things as witches. It would be madness to identify the property of *being accused of witchcraft* (or any property like *being a woman who threatens patriarchy in such-and-such a way*, and so on) with the property of *being a witch*. By the same token, when we employ folk psychology we "accuse" people of having beliefs and desires and so forth, and it may well be that there is something distinctive going on in their brains when we do this, but nevertheless (the ontological eliminativist thinks) it is a mistake to identify these complicated brain states with beliefs, desires, and so forth.

This solution to the challenge of cognitive suicide may encourage one to say something similar in response to the challenge of fictionalist suicide. The ontological eliminativist about psychological entities need not deny that when there is a paradigm instance (according to folk psychology) of someone's believing that *p*, there is something distinctive going on in the subject's brain: something that is different from what is going on in the brain of a person who is a paradigm instance (according to folk psychology) of someone's treating *p* as a make-believe. In other words, the ontological eliminativist need not hold that folk psychology is a *wild and whimsical* false theory utterly disconnected from what's going on in the world and in people's brains. Folk

[13] Quine writes: "Is physicalism a repudiation of mental objects after all, or a theory of them? Does it repudiate the mental state of pain or anger in favor of its physical concomitant, or does it identify the mental state with a state of the physical organism (and so a state of the physical organism with the mental state)?" (Quine 1960: 265). He thinks the question that forces the choice is misguided.

[14] I've used this example before in Joyce (2001: 96; 2011b: 531–3).

psychology probably imperfectly tracks real brain events; it's just that the conceptual apparati that folk psychology employs (beliefs, desires, and so on) are so removed from reality that it is best to deny that these entities exist. Perhaps another analogy would help. Phlogiston theory was not a wild and whimsical idea either; it was a solid scientific theory—but a false one. Paradigm instances of phlogiston escaping (as identified by the theory) really did map onto real phenomena: namely, instances of oxygen being consumed through combustion.

What then of psychological fictionalism? When there is a paradigm instance (according to folk psychology) of someone's believing that *p*, then let's just say that this person is in brain state B; and when there is a paradigm instance (according to folk psychology) of someone's make-believing that *p*, then let's say that this person is in brain state M. Psychological fictionalism can, then, be translated into claims that are acceptable to an ontological eliminativist. Instead of recommending that we make-believe rather than believe propositions of the form "Mary believes that *p*," the fictionalist recommends that we adopt brain state M rather than brain state B toward such propositions. Instead of analysing propositions of the form "Mary believes that *p*" as "According to folk psychology, Mary believes that *p*," and then claiming that the latter may be truly believed, the fictionalist can say that instead of taking brain state B toward the proposition "Mary believes that *p*," we should take brain state B toward the proposition "According to folk psychology, Mary believes that *p*."[15]

The fact that psychological fictionalist proposals can in principle be imperfectly redescribed as claims about brain states does mean that something "in the spirit" of psychological fictionalism may be true or reasonable. However, inasmuch as the thesis of fictionalism is essentially characterized with reference to folk-psychological entities, then if the brain states in question cannot be identified with the entities of folk psychology (as the ontological eliminativist insists they cannot be), then psychological fictionalism simply cannot be literally correct. For example, if one denies the existence of make-believe, then one simply cannot propose that a person should take an attitude literally of make-believe toward X—not, at least, if the recommendation is supposed to be one that could be complied with. If one denies the existence of belief, then one cannot propose that a discourse consists literally of assertions containing a tacit story operator—not if to assert something is to express a belief in that thing. Strictly speaking, then, one might continue to insist that psychological fictionalism, as it is usually

[15] Wallace writes: "[I]t is not that the Mental Fictionalist thinks that absolutely *nothing* is happening in the world when someone claims to be pretending she is Superman or that she is drinking tea with her imaginary friend. There *is* cognitive activity when these things are seemingly going on. It is just that our ordinary, everyday pretense-sincere avowal talk is wildly disparate from the cognitive activity that *is* going on—so much so, that terms such as 'pretense,' 'make-believe,' and 'sincere avowal' fail to pick out any activity or process that's actually in the world. However, the advantage of Mental Fictionalism is that it allows one to still *talk as if* these terms do pick out something, even though this 'talk as if' will ultimately (and strictly speaking) get cashed out in terms of some sort of complicated cognitive activity that is unsuitable as a legitimate element of Folk Psychology" (unpublished).

articulated, does indeed commit suicide, even if some other theory in a similar "spirit" remains viable.

But despite its suicidal tendencies, psychological fictionalism can be in some measure resurrected. As we have seen, the ontological eliminativist may allow that the analyses and proposals of fictionalism can be taken to roughly concern real brain states. Lacking a proper vocabulary for discussing these brain states with the appropriate specificity, however, one seems justified in using the best shorthand available for denoting them: namely, the terms of folk psychology. The fictionalist is not proposing that we identify, say, *brain state M* with *acts of make-believe* (for that would be to abandon ontological eliminativism), but rather that we use the erroneous terms as a kind of imperfect proxy for complicated phenomena that we might otherwise struggle to describe. In a similar way, an anthropologist studying a foreign culture may find it convenient simply to speak of those actions and objects that are "tapu" (something in which he disbelieves), rather than laboriously referring to "what the people in this tribe tend to classify as tapu."

When someone in this manner says something that he himself judges to be false, for the sake of convenience, is this a kind of fictive act? Not necessarily. If I threaten to shoot someone unless he utters the sentence "1 + 2 = 5," then his prudent decision to comply does not really count as a fictive act. Such an act of linguistic convenience wouldn't involve the speaker *immersing* himself in any fiction; it doesn't involve him allowing falsehoods a role in deliberations. Moreover, we need not classify his utterance as one that we would ordinarily think of as ontologically committing but which is not ontologically committing, for I think we are all pretty clear that what a person says with a gun to his head needn't be anything he really means.

Similarly, when someone wishes to recommend adopting brain state B (or M, and so on), but lacks the vocabulary for doing so and thus speaks in terms of "belief" (or "make-believe," and so on) this act of linguistic convenience is not itself a fictive performance—not, at least, in the sense in which we are interested. Such a speaker is not herself following any fictionalist advice; she isn't conforming to any recommendation to make-believe that *p*, or to believe that according to *F*, *p*. Thus the psychological fictionalist may be justified in articulating the theory in exactly the way it is usually articulated: using terms like "belief," "make-believe," "acceptance," and "assertion." This will not be to speak literally; the fictionalist is adopting a language of convenience when she uses these terms to present her favored theory. But in using this familiar language the fictionalist is not presupposing the accuracy of any of the conceptual machinery of folk psychology. There are literal recommendations and analyses standing behind these specious terms, just not a convenient language with which to express them.

References

Ainslie, G. (1975). "Impulsiveness and impulse control." *Psychological Bulletin* 82: 463–96.

Alexander, R. (1987). *The Biology of Moral Systems*. Hawthorne, NY: Aldine de Gruyter.

Alicke, M. (1985). "Global self-evaluation as determined by the desirability and controllability of trait adjectives." *Journal of Personality and Social Psychology* 49: 1621–30.

Alston, W. (1996). "Belief, acceptance, and religious faith." In *Faith, Freedom, and Rationality*, ed. J. Jordan and D. Howard-Snyder: 3–27. Lanham, MD: Rowman and Littlefield.

Anscombe, G. E. M. (1958). "Modern moral philosophy." *Philosophy* 33: 1–19.

Árdal, P. (1966). *Passion and Value in Hume's Treatise*. Edinburgh: University of Edinburgh Press.

Ariew, A. (1996). "Innateness and canalization." *Philosophy of Science* suppl. vol. 63: S19–27.

Ariew, A. (1999). "Innateness is canalization: In defense of a developmental account of innateness." In *Where Biology Meets Psychology: Philosophical Essays*, ed. V. Hardcastle: 117–38. Cambridge, MA: MIT Press.

Aulas, J. and Rosner, I. (2003). "Efficacy of a non blind placebo prescription." *Encephale* 29: 68–71.

Austin, J. L. (1961). "Performative utterances." In *Philosophical Papers*, ed. J. Urmson and G. J. Warnock: 220–39. Oxford: Clarendon Press.

Austin, J. L. (1962). *How to Do Things with Words.* Oxford: Oxford University Press.

Austin, J. L. (1963). "Performative-constative." In *Philosophy and Ordinary Language*, ed. C. Canton: 22–54. Urbana, IL: University of Illinois Press.

Ayala, F. (2010). "The difference of being human: Morality." *Proceedings of the National Academy of Sciences of the USA* 107: 9015–22.

Ayer, A. J. ([1936] 1971). *Language, Truth and Logic*. New York: Penguin Books.

Ayer, A. J. (1980). *Hume*. New York: Hill and Wang.

Baldwin, T. (2002). "The three phases of intuitionism." In *Ethical Intuitionism: Re-evaluations*, ed. P. Stratton-Lake: 92–112. Oxford: Oxford University Press.

Baron-Cohen, S. (1987). "Autism and symbolic play." *British Journal of Developmental Psychology* 5: 139–48.

Bechara, A., Tranel, D., and Damasio, H. (2000). "Characterization of the decision-making deficit of patients with ventromedial prefrontal cortex lesions." *Brain* 123: 2189–202.

Blackburn, S. (1984). *Spreading the Word*. Oxford: Clarendon Press.

Blackburn, S. (1993). *Essays in Quasi-Realism.* Oxford: Oxford University Press.

Blackburn, S. (1995). "The flight to reality." In *Virtues and Reasons: Philippa Foot and Moral Theory*, ed. R. Hursthouse, G. Lawrence, and W. Quinn: 35–56. Oxford: Clarendon Press.

Blackburn, S. (1998). *Ruling Passions*. Oxford: Oxford University Press.

Blair, J., Mitchell, D., and Blair, K. (2005). *The Psychopath: Emotion and the Brain*. Oxford: Blackwell.

Boehm, C. (2012). *Moral Origins: The Evolution of Virtue, Altruism, and Shame*. New York: Basic Books.

Boghossian, P. (1990). "The status of content." *Philosophical Review* 99: 157–84.

Borgida, E. and Nisbett, R. (1977). "The differential impact of abstract vs. concrete information on decisions." *Journal of Applied Social Psychology* 7: 258–71.

Brandom, R. (1994). *Making it Explicit*. Cambridge, MA: Harvard University Press.

Brandt, R. (1955). "The definition of an 'ideal observer' theory in ethics." *Philosophy and Phenomenological Research* 15: 407–13.

Bratman, M. (1992). "Practical reasoning and acceptance in a context." *Mind* 101: 1–15.

Bricke, J. (1996). *Mind and Morality: An Examination of Hume's Moral Psychology.* Oxford: Oxford University Press.

Brink, D. (1989). *Moral Realism and the Foundations of Ethics*. Cambridge: Cambridge University Press.

Broad, C. D. (1930). *Five Types of Ethical Theory*. New York: Harcourt, Brace and Co.

Brogaard, B. and Salerno, J. (2013). "Remarks on counterpossibles." *Synthese* 190: 639–69.

Brosnan, K. (2011). "Do the evolutionary origins of our moral beliefs undermine moral knowledge?" *Biology and Philosophy* 26: 51–64.

Budd, M. (1985). *Music and the Emotions*. London: Routledge and Kegan Paul.

Buller, D. (2006). *Adapting Minds*. Cambridge, MA: MIT Press.

Burgess, J. ([1978] 2010). "Against ethics." In *A World Without Values: Essays on John Mackie's Moral Error Theory*, ed. R. Joyce and S. Kirchin: 1–15. Dordrecht: Springer.

Burgess, J. (1983). "Why I am not a nominalist." *Notre Dame Journal of Formal Logic* 24: 93–105.

Burgess, J. and Rosen, G. (1997). *A Subject with No Object*. Oxford: Clarendon Press.

Butterworth, B. (1999). *What Counts? How Every Brain is Hardwired for Math*. New York: Free Press.

Camus, A. (1951). *L'Homme révolté*. Paris: Gallimard.

Carlsmith, J. and Gross, A. (1969). "Some effects of guilt on compliance." *Journal of Personality and Social Psychology* 11: 232–9.

Carnap, R. ([1928] 1967). *The Logical Structure of the World*. London: Routledge and Kegan Paul.

Carnap, R. (1935). *Philosophy and Logical Syntax*. London: Kegan Paul, Trench, Trubner and Co. Ltd.

Carnap, R. (1950a). *Logical Foundations of Probability*. Chicago, IL: University of Chicago Press.

Carnap, R. (1950b). "Empiricism, semantics, and ontology." Reprinted in his *Meaning and Necessity*: 205–21. Chicago, IL: University of Chicago Press.

Carruthers, P. and James, S. (2008). "Evolution and the possibility of moral realism." *Philosophy and Phenomenological Research* 77: 237–44.

Carson, T. (1984). *The Status of Morality*. Boston, MA: Reidel.

Carson, T. (1989). "Could ideal observers disagree?: A reply to Taliaferro." *Philosophy and Phenomenological Research* 50: 115–24.

Casati, R. and Tappolet, C. (eds.). (1998). *European Review of Philosophy 3: Response-dependence*. Stanford, CA: CSLI Publications.

Chomsky, N. (1967). "Recent contributions to the theory of innate ideas." *Synthese* 17: 2–11.

Chomsky, N. (1990). "On the nature, use, and acquisition of language." In *Mind and Cognition*, ed. W. Lycan: 627–46. Oxford: Blackwell.

Churchland, Patricia (1981). "Is determinism self-refuting?" *Mind* 90: 99–101.

Churchland, Patricia (1986). *Neurophilosophy: Toward a Unified Science of the Mind/Brain*. Cambridge, MA: MIT Press.

Churchland, Patricia (2011). *Braintrust: What Neuroscience Tells Us about Morality*. Princeton, NJ: Princeton University Press.

Churchland, Paul (1979). *Scientific Realism and the Plasticity of Mind*. New York: Cambridge University Press.

Churchland, Paul (1981). "Eliminative materialism and the propositional attitudes." *Journal of Philosophy* 78: 67–90.

Ciaramelli, E., Muccioli, M., Làdavas, E., and di Pellegrino, G. (2007). "Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex." *Social Cognitive and Affective Neuroscience* 2: 84–92.

Clarke-Doane, J. (2012). "Morality and mathematics: The evolutionary challenge." *Ethics* 122: 313–40.

Cleckley, H. (1941). *The Mask of Sanity: An Attempt to Reinterpret the So-called Psychopathic Personality*. St Louis, MO: CV Mosby Company.

Cling, A. (1989). "Eliminative materialism and self-referential inconsistency." *Philosophical Studies* 56: 53–75.

Cohen, L. (1992). *An Essay on Belief and Acceptance*. Oxford: Clarendon Press.

Cohen, R. (1997). "Is Hume a noncognitivist in the motivation argument?" *Philosophical Studies* 85: 251–66.

Conee, E. and Feldman, R. (1998). "The generality problem for reliabilism." *Philosophical Studies* 89: 1–29.

Copp, D. (2001). "Realist-expressivism." *Social Philosophy and Policy* 18: 1–43.

Cosmides, L. and Tooby, J. (1992). "Cognitive adaptations for social exchange." In *The Adapted Mind*, ed. J. Barkow, L. Cosmides, and J. Tooby: 163–228. Oxford: Oxford University Press.

Cowie, F. (1999). *What's Within: Nativism Reconsidered*. Oxford: Oxford University Press.

Craig, E. (2000). "Hume on causality: Projectivist and realist?" In *The New Hume Debate*, ed. R. Read and K. Richman: 113–21. London: Routledge.

Currie, G. (1990). *The Nature of Fiction*. Cambridge: Cambridge University Press.

Currie, G. (1995). "Visual imagery as the simulation of vision." *Mind & Language* 10: 25–44.

D'Arms, J. and Jacobson, D. (2006). "Sensibility theory and projectivism." In *The Oxford Handbook of Ethical Theory*, ed. D. Copp: 186–218. Oxford: Oxford University Press.

Dahaene, S. (1997). *The Number Sense: How the Mind Creates Mathematics*. Oxford: Oxford University Press.

Dancy, J. (1986). "Two conceptions of moral realism." *Proceedings of the Aristotelian Society* suppl. vol. 60: 167–87.

Darwall, S. (2000). "Normativity and projection in Hobbes's *Leviathan*." *Philosophical Review* 109: 313–47.

Darwin, C. (1877). "A biographical sketch of an infant." *Mind* 2: 285–94.

Darwin, C. ([1879] 2004). *The Descent of Man*. London: Penguin Books.

Davies, M. and Stone, T. (eds.). (1995a). *Folk Psychology: The Theory of Mind Debate*. Oxford: Blackwell.

Davies, M. and Stone, T. (eds.). (1995b). *Mental Simulation: Evaluations and Applications*. Oxford: Blackwell.

De Waal, F. B. M. (1992). "The chimpanzee's sense of social regularity and its relation to the human sense of justice." In *The Sense of Justice: Biological Foundations of Law*, ed. R. Masters and M. Gruter: 241–55. Newbury Park, CA: Sage Publications.

De Waal, F. B. M. (1996). *Good Natured: The Origins of Right and Wrong in Primates and Other Animals*. Cambridge, MA: Harvard University Press.

De Waal, F. B. M. (2006). *Primates and Philosophers*. Princeton, NJ: Princeton University Press.

Dennett, D. (1987). *The Intentional Stance*. Cambridge, MA: MIT Press.

Dennett, D. (1988). "Quining qualia." In *Consciousness and Contemporary Science*, ed. A. Marcel and E. Bisiach: 42–77. Oxford: Oxford University Press.

Dennett, D. (1995). *Darwin's Dangerous Idea*. New York: Simon and Schuster.

Descartes, R. ([1641] 1970). "Letter to Abbe de Launey." In *Descartes: Philosophical Letters*, ed. A. Kenny: 109. Oxford: Clarendon Press.

Devitt, M. and Rey, G. (1991). "Transcending transcendentalism." *Pacific Philosophical Quarterly* 72: 87–100.

Doris, J. (2002). *Lack of Character: Personality and Moral Behavior*. Cambridge: Cambridge University Press.

Dummett, M. (1959). "Truth." *Proceedings of the Aristotelian Society* 59: 141–62.

Dummett, M. (1978). *Truth and Other Enigmas*. London: Duckworth.

Dummett, M. (1981). *Frege: Philosophy of Language*. Cambridge, MA: Harvard University Press.

Dummett, M. (1993). *The Seas of Language*. Oxford: Oxford University Press.

Dwyer, S. (2006). "How good is the linguistic analogy?" In *The Innate Mind: Culture and Cognition*, ed. P. Carruthers, S. Laurence, and S. Stich: 237–55. Oxford: Oxford University Press.

Dwyer, S. (2009). "Moral dumbfounding and the linguistic analogy: Methodological implications for the study of moral judgment." *Mind & Language* 24: 274–96.

Dwyer, S., Huebner, B., and Hauser, M. (2010). "The linguistic analogy: Motivations, results, speculations." *Topics in Cognitive Science* 2: 486–510.

Elster, J. (1984). *Ulysses and the Sirens*. Cambridge: Cambridge University Press.

Elster, J. (1985). "Weakness of will and the free-rider problem." *Economics and Philosophy* 1: 231–65.

Enoch, D. (2010). "The epistemological challenge to metanormative realism: How best to understand it, and how to cope with it." *Philosophical Studies* 148: 413–38.

Enzle, M., Hansen, R., and Lowe, C. (1975). "Humanizing the mixed-motive paradigm: Methodological implications from attribution theory." *Simulation and Games* 6: 151–65.

Fessler, D., Arguello, A., Mekdara, J., and Macias, R. (2003). "Disgust sensitivity and meat consumption: A test of an emotivist account of moral vegetarianism." *Appetite* 41: 31–41.

Firth, R. (1952). "Ethical absolutism and the ideal observer." *Philosophy and Phenomenological Research* 12: 317–45.

Firth, R. (1955). "Reply to Professor Brandt." *Philosophy and Phenomenological Research* 15: 414–21.

Flack, J. and de Waal, F. B. M. (2001). "'Any animal whatever': Darwinian building blocks of morality in monkeys and apes." In *Evolutionary Origins of Morality: Cross-disciplinary Perspectives*, ed. L. Katz: 1–29. Thorverton: Imprint Academic.

Flanagan, O. (1992). *Consciousness Reconsidered*. Cambridge, MA: MIT Press.

Flew, A. (1963). "On the interpretation of Hume." *Philosophy* 38: 178–81.

Foot, P. (1972). "Morality as a system of hypothetical imperatives." *Philosophical Review* 81: 305–16.

Frankish, K. (1998). "A matter of opinion." *Philosophical Psychology* 11: 423–42.

Freedman, J. (1970). "Transgression, compliance, and guilt." In *Altruism and Helping Behavior*, ed. J. Macaulay and L. Berkowitz: 155–61. New York: Academic Press.

Frege, G. ([1892] 1997). "On sense and reference." In *The Frege Reader*, ed. M. Beaney: 151–71. Oxford: Basil Blackwell.

Frege, G. ([1897] 1997). "Logic." In *The Frege Reader*, ed. M. Beaney: 227–50. Oxford: Basil Blackwell.

Freud, S. ([1927] 1987). "The future of an illusion." In *Civilization, Society and Religion: The Pelican Freud Library, Vol. 12.* J. Strachey (trans.). London: Penguin.

Friedrich, J. (1996). "On seeing oneself as less self-serving than others: The ultimate self-serving bias?" *Teaching of Psychology* 23: 107–9.

Galileo, G. ([1623] 1960). "The assayer." In *The Controversy on the Comets of 1618*, ed. S. Drake and C. O'Malley: 151–336. Philadelphia, PA: University of Pennsylvania Press.

Garcha, H. and Ettlinger, G. (1979). "Object sorting by chimpanzees and monkeys." *Cortex* 15: 213–24.

Garner, R. (1993). "Are convenient fictions harmful to your health?" *Philosophy East and West* 43: 87–106.

Garner, R. (1994). *Beyond Morality*. Philadelphia, PA: Temple University Press.

Garner, R. (2010). "Abolishing morality." In *A World Without Values*, ed. R. Joyce and S. Kirchin: 217–33. Dordrecht: Springer Press.

Gauthier, D. (1992). "Artificial virtues and the sensible knave." *Hume Studies* 18: 401–27.

Goldman, A. (1979). "What is justified belief?" In *Justification and Knowledge*, ed. G. Pappas: 1–23. Dordrecht: Reidel.

Goldman, A. (1988). "Strong and weak justification." *Philosophical Perspectives* 2: 51–69.

Goldman, A. (1992). "Empathy, mind, and morals." *Proceedings and Addresses of the American Philosophical Association* 66: 17–43. Reprinted in Davies and Stone (1995b).

Goldman, A. (1993). "The psychology of folk psychology." *Behavioral and Brain Science* 16: 15–28.

Goldman, A. (2008). "Immediate justification and process reliabilism." In *Epistemology: New Essays*, ed. Q. Smith: 63–82. Oxford: Oxford University Press.

Goodwin, G. and Darley, J. (2008). "The psychology of meta-ethics: Exploring objectivism." *Cognition* 106: 1339–66.

Goodwin, G. and Darley, J. (2012). "Why are some moral beliefs perceived to be more objective than others?" *Journal of Experimental Social Psychology* 48: 250–6.

Gould, S. J. and Lewontin, R. C. (1979). "The spandrels of San Marco and the Panglossion paradigm: A critique of the adaptationist programme." *Proceedings of the Royal Society B* 205: 581–98.

Greene, J. (2008). "The secret joke of Kant's soul." In *Moral Psychology, Vol. 3: The Neuroscience of Morality*, ed. W. Sinnott-Armstrong: 35–79. Cambridge, MA: MIT Press.

Greene, J. and Haidt, J. (2002). "How and where does moral judgment work?" *Trends in Cognitive Sciences* 6: 517–23.

Greene, J., Nystrom, L., Engell, A., Darley, J., and Cohen, J. (2004). "The neural bases of cognitive conflict and control in moral judgment." *Neuron* 44: 389–400.

Greene, J., Sommerville, R., Nystrom, L., Darley, J., and Cohen, J. (2001). "An fMRI investigation of emotional engagement in moral judgment." *Science* 293: 2105–8.

Grice, P. (1989). *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.
Griffiths, P. (2002). "What is innateness?" *Monist* 85: 70–85.
Haack, S. (1976). "The pragmatist theory of truth." *British Journal for the Philosophy of Science* 27: 231–49.
Hägerström, A. (1953). *Inquiries into the Nature of Law and Morals*. Stockholm: Almqvist and Wiksell.
Haidt, J. (2001). "The emotional dog and its rational tail: A social intuitionist approach to moral judgment." *Psychological Review* 108: 814–34.
Haidt, J. and Joseph, C. (2004). "Intuitive ethics: How innately prepared intuitions generate culturally variable virtues." *Daedalus* 133: 55–66.
Haji, I. (1998). *Moral Appraisability*. Oxford: Oxford University Press.
Haji, I. (2003). *Deontic Morality and Control.* Cambridge: Cambridge University Press.
Hamilton, W. (1964). "The genetical evolution of social behavior, I and II." *Journal of Theoretical Biology* 7: 1–52.
Hare, R. (1993). *Without Conscience: The Disturbing World of the Psychopaths Among Us*. New York: Pocket Books.
Hare, R. M. (1952). *The Language of Morals.* Oxford: Oxford University Press.
Hare, R. M. (1999). *Objective Prescriptions and Other Essays.* Oxford: Oxford University Press.
Harman, G. (1977). *The Nature of Morality: An Introduction to Ethics*. New York: Oxford University Press.
Harman, G. (1986). "Moral explanations of natural facts: Can moral claims be tested against moral reality?" *Southern Journal of Philosophy* suppl. vol. 24: 57–68.
Harman, G. (1999). "Moral philosophy meets social psychology: Virtue ethics and the fundamental attribution error." *Proceedings of the Aristotelian Society* 99: 315–31.
Harman, G. (2000). "The nonexistence of character traits." *Proceedings of the Aristotelian Society* 100: 223–6.
Harman, G. and Thomson, J. J. (1996). *Moral Relativism and Moral Objectivity.* Oxford: Blackwell.
Hauser, M. (2006). *Moral Minds*. New York: Harper Collins.
Hauser, M., Chomsky, N., and Fitch, W. T. (2002). "The faculty of language: What is it, who has it, and how did it evolve?" *Science* 298: 1569–79.
Heil, J. (2005). "Dispositions." *Synthese* 144: 343–56.
Heine, B. (1985). "The mountain people: Some notes on the Ik of north-eastern Uganda." *Africa: Journal of the International African Institute* 55: 3–16.
Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., and Gintis, H. (2004). *Foundations of Human Sociality*. New York: Oxford University Press.
Hinckfuss, I. (1987). "The moral society: Its structure and effects." *Discussion Papers in Environmental Philosophy* 16. Canberra: Philosophy Program (RSSS), Australian National University.
Hirstein, W. (2005). *Brain Fiction: Self-deception and the Riddle of Confabulation*. Cambridge, MA: MIT Press.
Hobbes, T. ([1642] 1983). *De Cive*. Oxford: Clarendon Press.
Horberg, E., Oveis, C., and Keltner, D. (2011). "Emotions as moral amplifiers: An appraisal tendency approach to the influences of distinct emotions upon moral judgment." *Emotion Review* 3: 237–44.

Horgan, T. and Tienson, J. (2002). "The intentionality of phenomenology and the phenomenology of intentionality." In *Philosophy of Mind: Classical and Contemporary Readings*, ed. D. Chalmers: 520–33. New York: Oxford University Press.

Horgan, T. and Timmons, M. (2008). "What does moral phenomenology tell us about moral objectivity?" *Social Philosophy and Policy* 25: 267–300.

Horgan, T. and Woodward, J. (1985). "Folk psychology is here to stay." *Philosophical Review* 94: 197–226.

Hornstein, H., Lakind, E., Frankel, E., and Manne, S. (1975). "Effects of knowledge about remote social events on prosocial behaviour, social conception, and mood." *Journal of Personality and Social Psychology* 32: 1038–46.

Huemer, M. (2005). *Ethical Intuitionism*. New York: Palgrave Macmillan.

Hume, D. ([1740] 1978). *A Treatise of Human Nature*. Oxford: Clarendon Press.

Hume, D. ([1742] 1996). "The sceptic." In *David Hume: Selected Essays*: 95–113. Oxford: Oxford University Press.

Hume, D. ([1751] 1998). *An Enquiry Concerning the Principles of Morals*. Oxford: Clarendon Press.

Hume, D. ([1757] 1996). "Of the standard of taste." In *David Hume: Selected Essays*: 133–54. Oxford: Oxford University Press.

Hume, D. ([1762] 1986). "A new letter to Hugh Blair from July 1762." *Mind* 95: 411–16.

Hursthouse, R. (1987). *Beginning Lives*. Oxford: Wiley–Blackwell.

Hursthouse, R. (1995). "Applying virtue ethics." In *Virtues and Reasons: Philippa Foot and Moral Theory*, ed. R. Hursthouse, G. Lawrence, and W. Quinn: 57–75. Oxford: Clarendon Press.

Hussain, N. (2004). "The return of moral fictionalism." *Philosophical Perspectives* 18: 149–87.

Irons, W. (1996). "Morality as an evolved adaptation." In *Investigating the Biological Foundations of Human Morality*, ed. J. Hurd: 1–34. Lewiston, NY: Edwin Mellen Press.

Jackson, F. (1998). *From Metaphysics to Ethics*. Oxford: Oxford University Press.

Jarrold, C., Boucher, J., and Smith, P. (1996). "Generativity deficits in pretend play in autism." *British Journal of Developmental Psychology* 14: 275–300.

Johnson-Laird, P. and Oatley, K. (2000). "Cognitive and social construction in emotions." In *Handbook of the Emotions*, ed. M. Lewis and J. Haviland-Jones: 341–52. New York: Guilford Press.

Johnston, M. (1989). "Dispositional theories of value." *Proceedings of the Aristotelian Society* suppl. vol. 63: 139–74.

Johnston, M. (1992). "How to speak of the colors." *Philosophical Studies* 68: 221–63.

Johnston, M. (1993). "Objectivity refigured: Pragmatism without verificationism." In *Reality, Representation, and Projection*, ed. J. Haldane and C. Wright: 85–130. Oxford: Oxford University Press.

Joyce, R. (2000). "Rational fear of monsters." *British Journal of Aesthetics* 40: 209–24.

Joyce, R. (2001). *The Myth of Morality*. Cambridge: Cambridge University Press.

Joyce, R. (2002). "Expressivism and motivation internalism." *Analysis* 62: 336–44.

Joyce, R. (2005). "Moral fictionalism." In *Fictionalism in Metaphysics*, ed. M. Kalderon: 287–313. Oxford: Oxford University Press.

Joyce, R. (2006a). *The Evolution of Morality*. Cambridge, MA: MIT Press.

Joyce, R. (2006b). "Is human morality innate?" In *The Innate Mind: Culture and Cognition*, ed. P. Carruthers, S. Laurence, and S. Stich: 257–79. Oxford: Oxford University Press.

Joyce, R. (2007a). "Moral anti-realism." *The Stanford Encyclopedia of Philosophy*, ed. E. Zalta. <http://plato.stanford.edu/archives/sum2015/entries/moral-anti-realism/>

Joyce, R. (2007b). "Morality, schmorality." In *Morality and Self-Interest*, ed. P. Bloomfield: 51–75. Oxford: Oxford University Press.

Joyce, R. (2008). "What neuroscience can and cannot contribute to metaethics." In *Moral Psychology, Volume 3: The Neuroscience of Morality: Emotion, Brain Disorders, and Development*, ed. W. Sinnott-Armstrong: 371–94. Cambridge, MA: MIT Press.

Joyce, R. (2009a). "Is moral projectivism empirically tractable?" *Ethical Theory and Moral Practice*: 12: 53–75.

Joyce, R. (2009b). "Review of Jesse Prinz's *The Emotional Construction of Morals*." *Mind* 118: 508–18.

Joyce, R. (2009c). "The skeptick's tale." *Philosophy and Phenomenological Research* 78: 213–21.

Joyce, R. (2010a). "Patterns of objectification." In *A World Without Values: Essays on John Mackie's Moral Error Theory*, ed. R. Joyce and S. Kirchin: 35–53. Dordrecht: Springer Press.

Joyce, R. (2010b). "Expressivism, motivation internalism, and Hume." In *Hume on Motivation and Virtue*, ed. C. Pigden: 30–56. London: Palgrave MacMillan.

Joyce, R. (2011a). "The accidental error theorist." In *Oxford Studies in Metaethics, Vol. 6*, ed. R. Shafer-Landau: 153–80. Oxford: Oxford University Press.

Joyce, R. (2011b). "The error in 'The error in the error theory.'" *Australasian Journal of Philosophy* 89: 519–34.

Joyce, R. (2012). "Metaethical pluralism: How both moral naturalism and moral skepticism may be permissible positions." In *Ethical Naturalism: Current Debates*, ed. S. Nuccetelli and G. Seay: 89–109. Cambridge: Cambridge University Press.

Joyce, R. and Kirchin, S. (eds.) (2010). *A World Without Values: Essays on John Mackie's Moral Error Theory*. Dordrecht: Springer Press.

Kahane, G. (2011). "Evolutionary debunking arguments." *Noûs* 45: 103–25.

Kalderon, M. (2005a). *Moral Fictionalism*. Oxford: Oxford University Press.

Kalderon, M. (ed.) (2005b). *Fictionalism in Metaphysics*. Oxford: Oxford University Press.

Kelly, D. (2011). *Yuck! The Nature and Moral Significance of Disgust*. Cambridge, MA: MIT Press.

Kelly, D. (2014). "Selective debunking arguments, folk psychology, and empirical moral psychology." In *Advances in Experimental Moral Psychology*, ed. J. Wright and H. Sarkissian: 130–47. New York: Continuum Press.

Kelly, D. and Stich, S. (2007). "Two theories about the cognitive architecture underlying morality." In *The Innate Mind: Foundations and the Future*, ed. P. Carruthers, S. Laurence, and S. Stich: 348–67. Oxford: Oxford University Press.

Kelly, D., Stich, S., Haley, K., Eng, S., and Fessler, D. (2007). "Harm, affect and the moral/conventional distinction." *Mind & Language* 22: 117–31.

Kennett, J. (2002). "Autism, empathy and moral agency." *Philosophical Quarterly* 52: 340–57.

Kennett, J. and Fine, C. (2008). "Internalism and the evidence from psychopaths and 'acquired sociopaths'" (plus comments and reply). In *Moral Psychology, Vol 3: The Neuroscience of Morality: Emotion, Brain Disorders, and Development*, ed. W. Sinnott-Armstrong: 173–225. Cambridge, MA: MIT Press.

Kenny, A. (1964). *Action, Emotion and Will*. London: Routledge and Kegan Paul.
Ketelaar, T. and Au, W. (2003). "The effects of feelings of guilt on the behavior of uncooperative individuals in repeated social bargaining games: An affect-as-information interpretation of the role of emotion in social interaction." *Cognition and Emotion* 17: 429–53.
Keynes, J. M. (1949). *Two Memoirs*. London: Rupert Hart-Davis.
Kirchin, S. (2003). "Ethical phenomenology and metaethics." *Ethical Theory and Moral Practice* 6: 241–64.
Kitcher, P. (2011). *The Ethical Project*. Cambridge, MA: Harvard University Press.
Kitcher, P. (2014). "Is a naturalized ethics possible?" *Behaviour* 151: 245–60.
Knight, J. (1994). "The mountain people as tribal mirror." *Anthropology Today* 10: 1–3.
Krebs, D. (2005). "The evolution of morality." In *The Handbook of Evolutionary Psychology*, ed. D. Buss: 747–71. NY: Wiley.
Kriegel, U. (2003). "Consciousness as sensory quality and as implicit self-awareness." *Phenomenology and the Cognitive Sciences* 2: 1–26.
Kriegel, U. (2008). "Moral phenomenology: Foundational issues." *Phenomenology and the Cognitive Sciences* 7: 1–19.
Lahav, R. (1992). "The amazing predictive power of folk psychology." *Australasian Journal of Philosophy* 70: 99–105.
Lamarque, P. (1981). "How can we fear and pity fictions?" *British Journal of Aesthetics* 21: 291–304.
Lewis, D. (1970). "How to define theoretical terms." *Journal of Philosophy* 67: 427–46.
Lewis, D. (1978). "Truth in fiction." *American Philosophical Quarterly* 15: 37–46.
Lewis, D. (1989). "Dispositional theories of value." *Proceedings of the Aristotelian Society* suppl. vol. 63: 113–37.
Loeb, D. (2007). "The argument from moral experience." *Ethical Theory and Moral Practice* 10: 469–84.
Loeb, D. (2010). "The argument from moral experience." In *A World Without Values: Essays on John Mackie's Moral Error Theory*, ed. R. Joyce and S. Kirchin: 101–18. Dordrecht: Springer.
Lycan, W. (1986). "Moral facts and moral knowledge." *Southern Journal of Philosophy* suppl. vol. 24: 79–94.
Lyle, J. and Hoffman, H. (1971). "Children's use of television and other media." In *Television and Social Behavior*, ed. J. Murray, E. Robinson, and G. Comstock. Rockville, MD: National Institutes of Health.
MacFarlane, J. (2005). "Making sense of relative truth." *Proceedings of the Aristotelian Society* 105: 321–39.
Machery, E. and Mallon, R. (2010). "The evolution of morality." In *The Moral Psychology Handbook*, ed. J. Doris, G. Harman, S. Nichols, J. Prinz, W. Sinnott-Armstrong, and S. Stich: 3–46. Oxford: Oxford University Press.
MacIntyre, A. (1981). *After Virtue*. Notre Dame, IN: University of Notre Dame Press.
Mackie, J. L. (1946). "A refutation of morals." *Australasian Journal of Psychology and Philosophy* 24: 77–90.
Mackie, J. L. (1973). *Truth, Probability and Paradox*. Oxford: Oxford University Press.
Mackie, J. L. (1977). *Ethics: Inventing Right and Wrong*. London: Penguin.
Mackie, J. L. (1980). *Hume's Moral Theory.* New York: Routledge.
Mackie, J. L. (1982). *The Miracle of Theism*. Oxford: Clarendon Press.

Malcolm, N. (1958). *Ludwig Wittgenstein: A Memoir.* Oxford: Oxford University Press.

Mameli, M. (2008). "On innateness: The clutter hypothesis and the cluster hypothesis." *Journal of Philosophy* 105: 719–37.

Mameli, M., and Bateson, P. (2007). "The innate and the acquired: Useful clusters or a residual distinction from folk biology?" *Developmental Psychobiology* 49: 818–31.

Mandelbaum, M. (1955). *The Phenomenology of Moral Experience*. Glencoe, IL: Free Press.

Mares, E. (1997). "Who's afraid of impossible worlds?" *Notre Dame Journal of Formal Logic* 38: 516–26.

Marks, J. (2013). *Ethics without Morals: In Defence of Amorality*. New York: Routledge.

Mason, K. (2010). "Debunking arguments and the genealogy of religion and morality." *Philosophy Compass* 5: 770–8.

McKeever, S. (2003). "Review of Joyce, *The Myth of Morality*." *Ethics* 114: 182–4.

McKenna, F., Stanier, R., and Lewis, C. (1991). "Factors underlying illusory self-assessment of driving skill in males and females." *Accident Analysis and Prevention* 23: 45–52.

Merleau-Ponty, M. (1945). *Phénoménologie de la Perception*. Paris: Gallimard.

Merritt, M. (2002). "Virtue ethics and situationist personality psychology." *Ethical Theory and Moral Practice* 3: 365–83.

Mikhail, J. (2008). "The poverty of the moral stimulus." In *Moral Psychology, Vol. 1: The Evolution of Morality: Adaptations and Innateness*, ed. W. Sinnott-Armstrong: 353–60. Cambridge, MA: MIT Press.

Mikhail, J. (2011). *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*. Cambridge: Cambridge University Press.

Mill, J. S. (1861). "Utilitarianism, chapter 3." In *Fraser's Magazine*, November, 525–34.

Miller, G. (2007). "Sexual selection for moral virtues." *Quarterly Review of Biology* 82: 97–121.

Milo, R. (1995). "Contractarian constructivism." *Journal of Philosophy* 92: 181–204.

Moeller, H.-G. (2009). *The Moral Fool: A Case for Amorality*. New York: Columbia University Press.

Moll, J., de Oliveira-Souza, R., Eslinger, P., Bramati, I., Mourão-Miranda, J., Andreiuolo, P., and Pessoa, L. (2002). "The neural correlates of moral sensitivity: A functional magnetic resonance imaging investigation of basic moral emotions." *Journal of Neuroscience* 22: 2730–6.

Moore, G. E. (1942). "A reply to my critics." In *The Philosophy of G. E. Moore*, ed. P. Schilpp: 535–677. Evanston, IL: Northwestern University.

Mormann, T. (2007). "Carnap's logical empiricism, values, and American pragmatism." *Journal for the General Philosophy of Science* 38: 127–46.

Nesse, R. (2007). "Runaway social selection for displays of partner value and altruism." *Biological Theory* 2: 143–55.

Newman, J. (1981). "The fictionalist analysis of some moral concepts." *Metaphilosophy* 12: 47–56.

Newmark, C., Frerking, R., Cook, L., and Newmark, L. (1973). "Endorsement of Ellis' irrational beliefs as a function of psychopathology." *Journal of Clinical Psychology* 29: 300–2.

Nichols, S. (2004). *Sentimental Rules: On the Natural Foundations of Moral Judgment*. New York: Oxford University Press.

Nichols, S. (2005). "Innateness and moral psychology." In *The Innate Mind: Structure and Contents*, ed. P. Carruthers, S. Laurence, and S. Stich: 353–430. Oxford: Oxford University Press.

Nichols, S. and Folds-Bennett, T. (2003). "Are children moral objectivists?" *Cognition* 90: 23–32.

Nichols, S. and Stich, S. (2003). *Mindreading*. New York: Oxford University Press.

Nisbett, R. and Ross, L. (1980). *Human Inference: Strategies and Shortcomings of Social Judgment*. Englewood Cliffs, NJ: Prentice-Hall.

Noë, R. (2001). "Biological markets: Partner choice as the driving force behind the evolution of mutualisms." In *Economics in Nature: Social Dilemmas, Mate Choice, and Biological Markets*, ed. R. Noë, J. van Hoof, and P. Hammerstein: 93–118. Cambridge: Cambridge University Press.

Nolan, D., Restall, G., and West, C. (2005). "Moral fictionalism versus the rest." *Australasian Journal of Philosophy* 83: 307–30.

Nucci, L. (2001). *Education in the Moral Domain*. Cambridge: Cambridge University Press.

O'Regan, K. and Noë, A. (2001). "What it is like to see: A sensorimotor theory of perceptual experience." *Synthese* 129: 79–103.

Ogden, C. K. and Richards, I. A. (1923). *The Meaning of Meaning*. London: Routledge and Kegan Paul.

Park, L. and Covi, U. (1965). "Nonblind placebo trial." *Archives of General Psychiatry* 12: 336–45.

Peacocke, C. (1983). *Sense and Content*. Oxford: Oxford University Press.

Peacocke, C. (1999). "Conscious attitudes, attention, and self-knowledge." In *Knowing Our Own Minds*, ed. C. Wright, B. Smith, and C. Macdonald: 63–98. Oxford: Oxford University Press.

Peirce, C. S. (1934). "Belief and judgment." In *Collected Papers, Vol. 5*, ed. C. Hartshorne and P. Weiss. Cambridge, MA: Harvard University Press.

Pettit, P. (1991). "Realism and response-dependence." *Mind* 100: 587–626.

Pitt, D. (2004). "The phenomenology of cognition; or what is it like to think that *p*?" *Philosophy and Phenomenological Research* 69: 1–36.

Price, A. (1992). "Three types of projectivism." In *Psychoanalysis: Mind and Art*, ed. J. Hopkins and A. Savile: 110–28. Oxford: Blackwell.

Price, H. (1988). *Facts and the Function of Truth*. Oxford: Blackwell.

Prinz, J. (2007). *The Emotional Construction of Morals*. Oxford: Oxford University Press.

Prinz, J. (2008a). "Is morality innate?" In *Moral Psychology, Vol. 1: The Evolution of Morality: Adaptations and Innateness*, ed. W. Sinnott-Armstrong: 367–406. Cambridge, MA: MIT Press.

Prinz, J. (2008b). "Resisting the linguistic analogy: A commentary on Hauser, Young, and Cushman." In *Moral Psychology, Vol. 2: The Cognitive Science of Morality: Intuition and Diversity*, ed. W. Sinnott-Armstrong: 157–70. Cambridge, MA: MIT Press.

Prinz, J. (2009). "Against moral nativism." In *Stich and his Critics*, ed. D. Murphy and M. Bishop: 167–89. Malden: Blackwell.

Prinz, J. (2014). "Where do morals come from? A plea for a cultural approach." In *Empirically Informed Ethics: Morality between Facts and Norms*, ed. M. Christen, C. van Schaik, J. Fischer, M. Huppenbauer, and C. Tanner: 99–116. Dordrecht: Springer.

Pronin, E., Lin, D., and Ross, L. (2002). "The bias blind spot: Perceptions of bias in self versus others." *Personality and Social Psychology Bulletin* 28: 369–81.

Putnam, H. (1971). *Philosophy of Logic*. New York: Harper and Row.

Quine, W. V. (1960). *Word and Object*. Cambridge, MA: MIT Press.
Quine, W. V. (1961). *From a Logical Point of View*. Cambridge, MA: Harvard University Press.
Quine, W. V. (1975). "On empirically equivalent systems of the world." *Erkenntnis* 9: 313–28.
Quine, W. V. (1981). *Theories and Things*. Cambridge, MA: Harvard University Press.
Quine, W. V. (1986). "Reply to Roger F. Gibson, Jr." In *The Philosophy of W. V. Quine*, ed. L. Hahn and P. Schilpp: 684–5. La Salle, IL: Open Court.
Quine, W. V. (1989). "Three indeterminacies." In *Perspectives on Quine*, ed. R. Barrett and R. Gibson: 1–16. Oxford: Basil Blackwell.
Quine, W. V. (1990). *The Pursuit of Truth*. Cambridge, MA: Harvard University Press.
Ramsey, W., Stich, S., and Garon, J. (1990). "Connectionism, eliminativism, and the future of folk psychology." In *Philosophical Perspectives: Action Theory and Philosophy of Mind*, ed. J. Tomberlin: 499–533. Atascadero, CA: Ridgeview.
Reppert, V. (1992). "Eliminative materialism, cognitive suicide, and begging the question." *Metaphilosophy* 23: 378–92.
Restall, G. (1997). "Ways things can't be." *Notre Dame Journal of Formal Logic* 38: 583–96.
Ridge, M. (2006). "Sincerity and expressivism." *Philosophical Studies* 131: 487–510.
Robinson, D. (2009). "Moral functionalism, ethical quasi-relativism, and the Canberra Plan." In *Conceptual Analysis and Philosophical Naturalism*, ed. D. Braddon-Mitchell and R. Nola: 315–48. Cambridge, MA: MIT Press.
Roorda, J. (1997). "Fallibilism, ambivalence, and belief." *Journal of Philosophy* 94: 126–55.
Rorty, R. (1970). "In defense of eliminative materialism." *Review of Metaphysics* 24: 112–21.
Rosen, G. (1994). "Objectivity and modern idealism: What is the question?" In *Philosophy in Mind*, ed. M. Michael and J. O'Leary-Hawthorne: 277–319. Dordrecht: Kluwer.
Ross, L., Lepper, M., and Hubbard, M. (1975). "Perseverance in self-perception and social perception: Biased attributional processes in the debriefing paradigm." *Journal of Personality and Social Psychology* 32: 880–92.
Ruse, M. (1986). *Taking Darwin Seriously*. Oxford: Basil Blackwell.
Ruse, M. (2006). "Is Darwinian metaethics possible (and if it is, is it well-taken)?" In *Evolutionary Ethics and Contemporary Biology*, ed. G. Boniolo and G. de Anna: 13–26. Cambridge: Cambridge University Press.
Ruse, M. (2009). "Evolution and ethics: The sociobiological approach." In *Philosophy After Darwin*, ed. M. Ruse: 489–511. Princeton, NJ: Princeton University Press.
Ruskin, J. (1856). *Modern Painters, Vol. 3*. Excerpt in *Selections from the Work of John Ruskin*. 1908. Boston, MA: Houghton Mifflin.
Russell, B. (1910). "Pragmatism." In his *Philosophical Essays*: 87–126. London: Longmans, Green, and Co.
Ryle, G. (1949). *The Concept of Mind*. London: Hutchinson.
Sainsbury, R. M. (1998). "Projections and relations." *Monist* 81: 133–60.
Sarkissian, H., Park, J., Tien, D., Wright, J. C., and Knobe, J. (2011). "Folk moral relativism." *Mind & Language* 26: 482–505.
Sayre-McCord, G. (1986). "The many moral realisms." *Southern Journal of Philosophy* suppl. vol. 24: 1–22.
Scanlon, T. (1998). *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.
Schafer, K. (2008). "Practical reasoning and practical reasons in Hume." *Hume Studies* 34: 189–208.

Schelling, T. (1980). "The intimate contest for self-command." *The Public Interest* 60: 94–118.
Schilpp, P. (ed.) (1963). *The Philosophy of Rudolf Carnap*. La Salle, IL: Open Court.
Schino, G., and Aureli, F. (2010). "The relative roles of kinship and reciprocity in explaining primate altruism." *Ecology Letters* 13: 45–50.
Schmidtz, D. (1994). "Choosing ends." *Ethics* 104: 226–51.
Schnall, S., Benton, J., and Harvey, S. (2008). "With a clean conscience: Cleanliness reduces the severity of moral judgment." *Psychological Science* 19: 1219–22.
Schnall, S., Haidt, J., Clore, G., and Jordan, A. (2008). "Disgust as embodied moral judgment." *Personality and Social Psychology Bulletin* 34: 1096–109.
Schnider, A. (2008). *The Confabulating Mind: How the Brain Creates Reality*. Oxford: Oxford University Press.
Searle, J. R. (1969). *Speech Acts.* Cambridge: Cambridge University Press.
Searle, J. R. (1975). "The logical status of fictional discourse." *New Literary History* 6: 319–32.
Searle, J. R. (1983). *Intentionality*. New York: Cambridge University Press.
Sellars, W. (1956). "Empiricism and the philosophy of mind." In *Minnesota Studies in the Philosophy of Science, Volume I: The Foundations of Science and the Concepts of Psychology and Psychoanalysis*, ed. H. Feigl and M. Scriven: 253–329. Minneapolis, MN: University of Minnesota Press.
Shafer-Landau, R. (2005). *Moral Realism.* Oxford: Oxford University Press.
Shafer-Landau, R. (2007). "Moral realism: Introduction." In *Foundation of Ethics*, ed. R. Shafer-Landau and T. Cuneo: 157–62. Oxford: Blackwell.
Siegel, S. (2006a). "Which properties are represented in perception?" In *Perceptual Experience*, ed. T. Gendler Szabo and J. Hawthorne: 481–503. Oxford: Oxford University Press.
Siegel, S. (2006b). "Subject and object in the contents of visual experience." *Philosophical Review* 115: 355–88.
Singer, P. (2005). "Ethics and intuitions," *Journal of Ethics* 9: 331–52.
Sinnott-Armstrong, W. (2006). *Moral Skepticisms*. Oxford: Oxford University Press.
Sinnott-Armstrong, W. (2008). "Is moral phenomenology unified?" *Phenomenology and the Cognitive Sciences* 7: 85–97.
Skarsaune, K. (2011). "Darwin and moral realism: Survival of the iffiest." *Philosophical Studies* 152: 229–43.
Slote, M. (1979). "Assertion and belief." In *Papers on Language and Logic*, ed. J. Dancy: 177–90. Keele: Keele University Library.
Small, D. and Lerner, J. (2008). "Emotional policy: Personal sadness and anger shape judgments about a welfare case." *Political Psychology* 29: 149–68.
Smetana, J. (1993). "Understanding of social rules." In *The Development of Social Cognition: The Child as Psychologist*, ed. M. Bennett: 111–41. New York: Guilford.
Smith, M. (1994). *The Moral Problem*. Oxford: Blackwell.
Smith, M. (2002). "Exploring the implications of the dispositional theory of value." *Philosophical Issues* 12: 329–47.
Smith, M. (2006). "Is that all there is?" *Journal of Ethics* 10: 75–106.
Snare, F. (1975). "The argument from motivation." *Mind* 84: 1–9.
Snare, F. (1991). *Morals, Motivation, and Convention.* Cambridge: Cambridge University Press.
Sobel, D. (1999). "Do the desires of rational agents converge?" *Analysis* 59: 137–47.
Sober, E. (1984). *The Nature of Selection*. Cambridge, MA: MIT Press.

Sober, E. (1988). "What is evolutionary altruism?" *Canadian Journal of Philosophy* suppl. vol. 14: 75–99.

Sober, E. (2000). "Psychological egoism." In *The Blackwell Guide to Ethical Theory*, ed. H. LaFollette: 129–48. Oxford: Blackwell.

Sober, E. (2009a). "Absence of evidence and evidence of absence: Evidential transitivity in connection with fossils, fishing, fine-tuning, and firing squads." *Philosophical Studies* 143: 63–90.

Sober, E. (2009b). "Parsimony arguments in science and philosophy: A test case for naturalism$_P$." *Proceedings and Addresses of the American Philosophical Association* 83: 117–55.

Sober, E. and Wilson, D. S. (1998). *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge, MA: Harvard University Press.

Sreenivasan, G. (2002). "Errors about errors: Virtue theory and trait attribution." *Mind* 111: 47–68.

Sripada, C. (2008). "Nativism and moral psychology: Three models of the innate structure that shapes the contents of moral norms." In *Moral Psychology, Vol. 1: The Evolution of Morality: Adaptations and Innateness*, ed. W. Sinnott-Armstrong: 319–43. Cambridge, MA: MIT Press.

Sripada, C., and Stich, S. (2006). "A framework for the psychology of norms." *In The Innate Mind: Culture and Cognition*, ed. P. Carruthers, S. Laurence, and S. Stich: 280–301. Oxford: Oxford University Press.

Stanley, J. (2001). "Hermeneutic fictionalism." *Midwest Studies in Philosophy* 25: 36–71.

Steen, F. and Owen, S. (2001). "Evolution's pedagogy: An adaptationist model of pretense and entertainment." *Journal of Cognition and Culture* 1: 289–321.

Sterelny, K. (2010). "Moral nativism: A sceptical response." *Mind & Language* 25: 279–97.

Stevenson, C. L. (1937). "The emotive meaning of ethical terms." *Mind* 46: 14–31.

Stevenson, C. L. (1963). *Facts and Values*. New Haven, CT: Yale University Press.

Stich, S. (1983). *From Folk Psychology to Cognitive Science: The Case Against Belief*. Cambridge, MA: MIT Press.

Stich, S. (2007). "Evolution, altruism, and cognitive architecture: A critique of Sober and Wilson's argument for psychological altruism." *Biology and Philosophy* 22: 267–81.

Stich, S. (2008). "Some questions about *The Evolution of Morality*." *Philosophy and Phenomenological Research* 77: 228–36.

Street, S. (2006). "A Darwinian dilemma for realist theories of value." *Philosophical Studies* 127: 109–66.

Street, S. (2008). "Reply to Copp: Naturalism, normativity, and the varieties of realism worth worrying about." *Philosophical Issues* 18: 207–28.

Stroud, B. (2000). "'Gilding or staining' the world with 'sentiments' and 'phantasms.'" In *The New Hume Debate*, ed. R. Read and K. Richman: 16–30. London: Routledge.

Sturgeon, N. (2008). "Hume's metaethics: Is Hume a moral noncognitivist?" In *A Companion to Hume*, ed. E. Radcliffe: 513–28. Oxford, Blackwell.

Svavarsdóttir, S. (2006). "How do moral judgments motivate?" In *Contemporary Debates in Moral Theory*, ed. J. Dreier: 163–81. Oxford: Blackwell.

Swain, M. (ed.) (1970). *Induction, Acceptance, and Rational Belief*. Dordrecht: Reidel.

Swanton, C. (2003). *Virtue Ethics: A Pluralistic View*. NY: Oxford University Press.

Taliaferro, C. (1988). "Relativising the ideal observer theory." *Philosophy and Phenomenological Research* 49: 123–38.

Tangney, J. and Dearing, R. (2002). *Shame and Guilt*. New York: Guilford.

Tangney, J., and Fischer, K. (eds.) (1995). *Self-conscious Emotions: The Psychology of Shame, Guilt, Embarrassment, and Pride*. New York: Guilford Press.

Tangney, J., Mashek, D., and Stuewig, J. (2007). "Working at the social-clinical-community-criminology interface: The GMU inmate study." *Journal of Social and Clinical Psychology* 26: 1–28.

Tangney, J., Stuewig, J., Malouf, E., and Youman, K. (2013). "Communicative functions of shame and guilt." In *Cooperation and its Evolution*, ed. K. Sterelny, R. Joyce, B. Calcott, and B. Fraser: 485–502. Cambridge, MA: MIT Press.

Taylor, K. (1994). "How not to refute eliminative materialism." *Philosophical Psychology* 7: 101–25.

Taylor, S. and Brown, J. (1988). "Illusion and well-being: A social psychological perspective on mental health." *Psychological Bulletin* 103: 193–210.

Timmons, M. (1999). *Morality Without Foundations.* Oxford: Oxford University Press.

Tolhurst, W. (1998). "Seemings." *American Philosophical Quarterly* 35: 293–302.

Tranel, D. (1994). "Acquired sociopathy: The development of sociopathic behavior following focal brain damage." *Progress in Experimental Personality and Psychopathology Research* 17: 285–311.

Turiel, E., Killen, M., and Helwig, C. (1987). "Morality: Its structure, functions, and vagaries." In *The Emergence of Morality in Young Children*, ed. J. Kagan and S. Lamb: 155–243. Chicago, IL: University of Chicago.

Turnbull, C. (1972). *The Mountain People*. New York: Simon and Schuster.

Tye, M. (1999). "Phenomenal consciousness: The explanatory gap as a cognitive illusion." *Mind* 108: 705–25.

Uttich, K., Tsai, G., and Lombrozo, T. (2014). "Exploring metaethical commitments: Moral objectivity and moral progress." In *Advances in Experimental Moral Psychology*, ed. H. Sarkissian and J. C. Wright: 188–208. London: Bloomsbury Academic.

Vaihinger, H. (1935). *The Philosophy of 'As If.'* C. K. Ogden (trans.), London: Routledge and Kegan Paul Ltd.

Valdesolo, P. and DeSteno, D. (2006). "Manipulations of emotional context shape moral judgment." *Psychological Science* 17: 476–7.

Van Fraassen, B. (1980). *The Scientific Image*. Oxford: Oxford University Press.

Vander Laan, D. (2004). "Counterpossibles and similarity." In *Lewisian Themes: The Philosophy of David K. Lewis*, ed. F. Jackson and G. Priest: 258–75. Oxford: Oxford University Press.

Vision, G. (1994). "Fiction and fictionalist reductions." *Pacific Philosophical Quarterly* 74: 150–74.

Wallace, M. (Unpublished manuscript). "Mental fictionalism."

Wallace, R. (2003). "Review of Joyce, *The Myth of Morality*." *Notre Dame Philosophical Reviews* 2003.11.04.

Walton, K. (1978). "Fearing fictions." *Journal of Philosophy* 75: 5–27.

Walton, K. (1990). *Mimesis and Make-Believe*. Cambridge, MA: Harvard University Press.

West, S., Griffin, A., and Gardner, A. (2007). "Social semantics: Altruism, cooperation, mutualism, strong reciprocity, and group selection." *European Society for Evolutionary Biology* 20: 415–32.

Wheatley, T. (2009). "Everyday confabulation." In *Confabulation: Views from Neuroscience, Psychiatry, Psychology, and Philosophy*, ed. W. Hirstein: 203–22. Oxford: Oxford University Press.

Wheatley, T. and Haidt, J. (2005). "Hypnotically induced disgust makes moral judgments more severe." *Psychological Science* 16: 780–4.

White, R. (2010). "You just believe that because…" *Philosophical Perspectives* 24: 573–615.

Wielenberg, E. (2010). "On the evolutionary debunking of morality." *Ethics* 120: 441–64.

Willett, J. (1957). *Brecht on Theatre*. NY: Hill and Wang.

Williams, B. (1966). "Consistency and realism." *Proceedings of the Aristotelian Society* suppl. vol. 40: 1–22.

Williams, B. (1981). "Internal and external reasons." In his *Moral Luck*: 101–13. Cambridge: Cambridge University Press.

Williamson, S., Harpur, T., and Hare, R. (1991). "Abnormal processing of affective words by psychopaths." *Psychophysiology* 28: 260–73.

Williamson, T. (2000). *Knowledge and its Limits*. Oxford: Oxford University Press.

Wilson, J. (1993). *The Moral Sense*. New York: Free Press.

Wittgenstein, L. (1965). "Lecture on ethics." *Philosophical Review* 74: 3–12.

Wong, D. (2006). *Natural Moralities: A Defense of Pluralistic Relativism*. Oxford: Oxford University Press.

Wray, K. (2001). "Collective belief and acceptance." *Synthese* 129: 319–33.

Wright, C. (1988). "Moral values, projections, and secondary qualities." *Proceedings of the Aristotelian Society* suppl. vol. 62: 1–26.

Wright, C. (1992). *Truth and Objectivity*. Cambridge, MA: Harvard University Press.

Wright, C. (1996). "Truth in ethics." In *Truth in Ethics*, ed. B. Hooker: 19–34. Oxford: Blackwell.

Zahavi, A. (1977). "The cost of honesty: Further remarks on the handicap principle." *Journal of Theoretical Biology* 67: 603–5.

Zangwill, N. (1990). "Quasi-quasi-realism." *Philosophy and Phenomenological Research* 50: 583–94.

Zangwill, N. (1992). "Quietism." *Midwest Studies in Philosophy* 17: 160–76.

Zhong, C., Bohns, V., and Gino, F. (2010). "Good lamps are the best police: Darkness increases dishonesty and self-interested behavior." *Psychological Science* 21: 311–14.

Zhong, C. and Liljenquist, K. (2006). "Washing away your sins: Threatened morality and physical cleansing." *Science* 313: 1451–2.

Index

abolitionism: *See* "eliminativism"
Anscombe, G.E.M. 1, 83, 192
Árdal, P. 20 n. 4
Austin, J.L. 27, 28, 35
Ayala, F. 110
Ayer, A.J. 20, 26, 97

Baldwin, T. 81 n
Blackburn, S. 18, 20, 46, 131 n, 168–9, 196, 197, 199, 202 n
Bloomfield, P. 41, 44–8, 49, 52
Boehm, C. 113 n
Bratman, M. 62
Bricke, J. 21 n. 6
Brink, D. 204, 210, 239
Brosnan, K. 144, 153, 155–6
Burgess, J. 51, 220
byproducts (vs. adapatations): *See* "spandrels"

Carnap, R. 26, 94–8, 99, 104
Chomsky, N. 117, 139
Churchland, Patricia 121, 240, 241, 251
Churchland, Paul 240, 241, 247, 251
Clarke-Doane, J. 150 n
Cleckley, H. 217
Cohen, R. 22 n
conservatism, epistemic 152, 157–8, 187–9, 194, 203
Copp, D. 28–30, 33–4, 36

Dancy, J. 203, 210
Darwin, C. 109, 110, 112, 122, 125, 127–30, 135, 136
Dennett, D. 240 n. 1
Descartes, R. 195
De Waal, F. 112–13, 116, 118 n
dispositions 23, 68, 71–6, 78, 80–2, 91, 93–4, 115, 135 n. 9, 183–5, 236, 246
Doris, J. 83
Dummett, M. 205

eliminativism (abolitionism) 11, 37, 58–9, 61, 66, 229–30, 232, 239, 240–3, 244, 247, 248–50, 251–5
Elster, J. 256
emotivism 18, 20–1, 26, 36, 39, 64, 198 n. 6
Enoch, D. 153
Euhemerus 142, 154
expressivism (*See also* "emotivism") 17–27, 30 n. 15, 33, 35–7, 39–40, 197, 200

Falk, W.D. 41 n. 1
Fessler, D. 20
Fine, C. 217
Firth, R. 76–80
Fitch, W.T. 117, 139
Frege, G. 33–4
Freud, S. 142, 143, 158, 185 n. 7

Galileo, G. 195
Garner, R. 47, 51, 229
Gauthier, D. 37 n. 24
Gould, S.J. 111
Greene, J. 20, 139 n, 145, 172
Grice, P. 29, 34

Haidt, J. 20, 139 n, 170–2
Hare, R.M. 33 n
Harman, G. 83, 85, 148, 166–7
Hauser, M. 117, 139
Heil, J. 75
Hobbes, T. 195, 230, 231
Hume, D. 5, 17, 18, 21–2, 26, 27, 35–40, 58, 168, 169–70, 177, 179 n, 184 n, 193, 195–8, 200, 209 n, 214, 219, 230–1, 234
Hursthouse, R. 84, 85, 86

ideal observer theory 17, 76–80, 193
innateness 9–10, 122–4, 132

Jackson, F. 82 n

Kahane, G. 145–6
Kelly, D. 170 n. 12, 171–2
Kennett, J. 217
Keynes, J.M. 100
Kitcher, P. 119–20

Lewis, D. 3, 32 n, 40 n. 27, 57, 91–4, 96, 97, 100, 104–5

MacFarlane, J. 30 n. 16
Machery, E. 140
MacIntyre, A. 100
Mackie, J.L. 4, 11, 36–7, 51, 65, 67–9, 70–2, 83, 137, 177–81, 183, 184 n, 186–94, 196, 204, 219–20, 229, 231, 236
Malcom, N. 27 n. 12
Mallon, R. 140
Mameli, M. 123
Mandelbaum, M. 164 n, 205

Mares, E. 148
Mason, K. 144
Mill, J.S. 122
Milo, R. 146 n. 4
Moore, G.E. 27, 100

naturalism, moral 5, 6, 7, 8, 89–94, 96, 97, 100–5, 120, 154–6, 165, 166, 193, 194 (vs. global naturalism 89–90)
Newton, I. 195
Nichols, S. 110–12, 117, 134, 135–6, 137–8, 139, 140
Nietzsche, F. 142
noncognitivism (*See also* "expressivism") 1–5, 11, 17–18, 21, 22 n, 25–6, 29–30, 35, 39 n, 41–2, 50–1, 64–5, 67–8, 89–90, 94–100, 104, 131 n, 140, 159–60, 168–74, 179, 182, 186, 193, 196–202, 209, 210 n, 215 n, 226–7, 243–5, 249, 251

Ogden C.K. 25

pathetic fallacy 178
prescriptivism 18, 95
Prinz, J. 20 n. 4, 73–6, 110, 115, 124, 132–3, 134–7, 139

quasi-realism 99, 131 n, 168–9
queerness, argument from 40, 67, 177, 186–9, 193–4
Quine, W.V. 58, 104–5, 220, 241, 253 n. 13

realism, moral 2, 4–5, 7, 8–9, 17, 36–7, 42, 90 n, 99, 131 n, 140, 144–5, 152, 159, 162–4, 167, 168, 182–8, 193, 196–8, 200, 203–4, 239
relativism 7, 36, 56, 57, 73–4, 166, 207, 227, 248–50
reliabilism 147, 149, 156
response dependence 70–82, 183–4
Richards, I.A. 25
Ridge, M. 23
Rosen, G. 206–8
Ruse, M. 144, 147–8, 159, 163–6
Ruskin, J. 178
Russell, B. 99–100, 162
Ryle, G. 242

Sainsbury, R.M. 196–7
Scanlon, T. 82
Searle, J.R. 20, 23
Sellars, W. 242 n. 4
Singer, P. 145, 146 n. 5
Sinnott-Armstrong, W. 90 n
Smith, M. 22, 82 n, 205, 217 n. 24
Snare, F. 26 n. 11
Sober, E. 126, 153
spandrels (byproducts) 9, 110–12, 113, 121, 124, 128, 134–5, 153
Sripada, C. 124
Sterelny, K. 124, 134, 135–6, 138, 139
Stevenson, C.L. 21 n. 5, 26
Stich, S. 124, 138 n. 13, 140, 159, 162–3, 167
Street, S. 9, 144–5, 150
Stroud, B. 179 n, 184 n
Sturgeon, N. 21 n. 6
Swanton, C. 84–5

Turnbull, C. 133

Vaihinger, H. 58, 63, 249 n
virtue ethics 68, 83–8

Wallace, M. 254 n
Walton, K. 232
Wheatley, T. 170–2
White, R. 157
Wielenberg, E. 153, 155
Wilson, J.Q. 238
Wittgenstein, L. 27 n. 12, 51 n
Wong, D. 105 n
Wright, C. 60 n. 25, 166, 205

Zangwill, N. 199–201